DOCUMENT RESUME

ED 473 713 EA 032 090

TITLE	No Child Left Behind: What Will It Take?
INSTITUTION	Thomas B. Fordham Foundation, Washington, DC.
PUB DATE	2002-02-00
NOTE	110p.; Papers presented at the Thomas B. Fordham Foundation Conference (Washington, DC, February 13, 2002).
AVAILABLE FROM	Thomas B. Fordham Foundation, 1627 K Street, N.W., Suite 600, Washington, DC 20006. Tel: 202-223-5452; Fax: 202-223-9226; e-mail: backtalk@edexcellence.net; Web site: http://www.edexcellence.net/index.html. For full text: http://www.edexcellence.net/NCLBconference/NCLBreport.pdf.
PUB TYPE	Collected Works - General (020) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC05 Plus Postage.
DESCRIPTORS	*Accountability; *Educational Policy; Elementary Secondary Education; *State Standards; Testing Programs
IDENTIFIERS	*No Child Left Behind Act 2001

ABSTRACT

This document contains seven policy papers and two comments that focus on those aspects of the No Child Left Behind Act involving state academic standards and testing programs; the intersection of state testing and national assessment; tracking of yearly progress; and accountability at the state, district, and school levels. Papers include: (1) "Multiple Choices: How Will States Fill in the Blanks in Their Testing Systems?" (Matthew Gandal); (2) "Using NAEP to Confirm State Test Results: Opportunities and Problems" (Mark D. Reckase); (3) "Adequate Yearly Progress: Results, Not Process" (Lisa Graham, Billie J. Orr, and Brian J. Jones); (4) "No Child Left Behind: Who Is Included in New Federal Requirements?" (Richard J. Wenning, Paul A. Herdman, and Nelson Smith); (5) "Aggregation and Accountability" (David Figlio); (6) "Comments" (Michael D. Casserly); (7) "Implementing Title I Standards, Assessments and Accountability: Lessons from the Past, Challenges for the Future" (Michael Cohen); (8) "What Might Go Wrong with the Accountability Measures of the 'No Child Left Behind Act'" (Dan Goldhaber); and (9) "Comments" (Abigail Thernstrom). (RT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 473 713

No Child Left Behind: What Will It Take?

Papers prepared for a conference sponsored by
The Thomas B. Fordham Foundation

February 2002

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

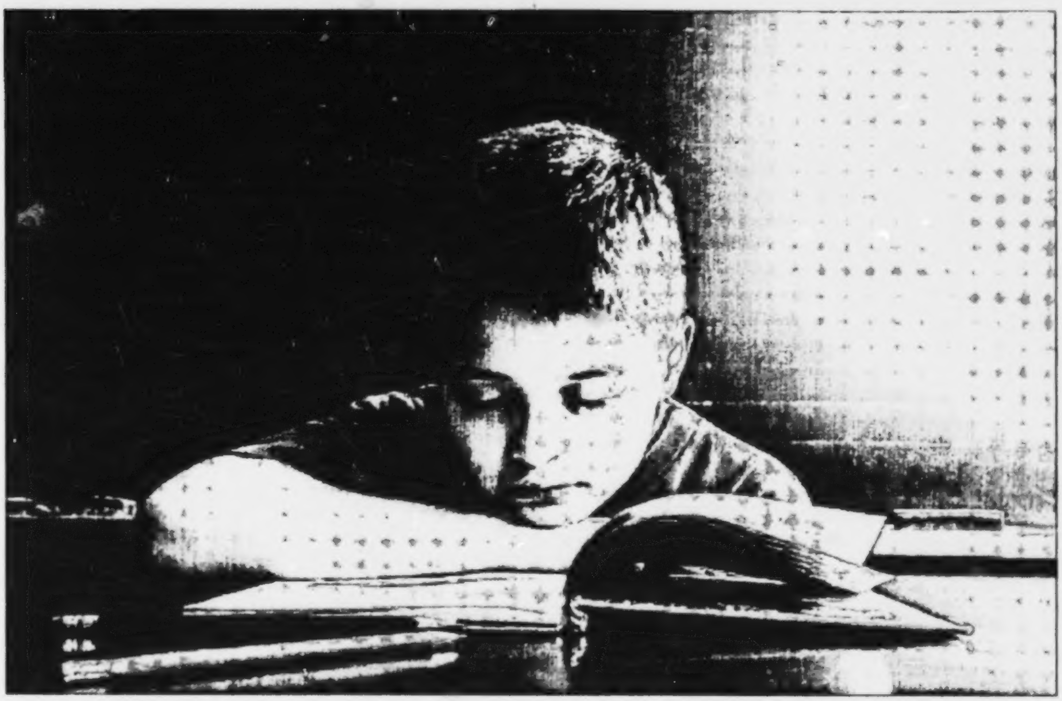
K. Amis

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

EA 032 090

BEST COPY AVAILABLE



No Child Left Behind: What Will It Take?

PAPERS PREPARED FOR A CONFERENCE SPONSORED
BY THE THOMAS B. FORDHAM FOUNDATION

THOMAS B.
FORDHAM
FOUNDATION
OUTSIDE THE BOX

February 2002

EA032090

Table of Contents

Foreword
Chester E. Finn, Jr. iv

State Testing Requirements and NAEP

Multiple Choices: How Will States Fill in the Blanks in their Testing Systems?
Matthew Gandal 1

Using NAEP to Confirm State Test Results: Opportunities and Problems
Mark D. Reckase 11

Adequate Yearly Progress Requirements

Adequate Yearly Progress: Results, not Process
Lisa Graham Keegan, Billie J. Orr & Brian J. Jones 21

No Child Left Behind: Who Is Included in New Federal Requirements?
Richard J. Wenning, Paul A. Herdman & Nelson Smith 35

Aggregation and Accountability
David Figlio 49

Comments
Michael D. Casserly 71

What can we learn from 1994 and what might go wrong in 2002?

Implementing Title I Standards, Assessments and Accountability: Lessons from the Past, Challenges for the Future
Michael Cohen 75

What Might Go Wrong with the Accountability Measures of the "No Child Left Behind Act"
Dan Goldhaber 89

Comments
Abigail Thernstrom 103

About the Contributors 109

Foreword

Chester E. Finn, Jr.

The No Child Left Behind Act is now just a month old, but it's already yowling and a lot of people are as nervous about it as new parents, unsure whether to feed it, hug it, put it to bed or spank it.

This is an enormous piece of legislation that possibly no human being has read from cover to cover. It spans dozens of programs and contains thousands of specific features. It ranges from Indian education to impact aid, from teacher quality to bilingual education, and on and on.

In the seven policy papers that follow, as in the conference that the Thomas B. Fordham Foundation hosted on February 13, the focus is on the implementation of the part of the act that got the most attention, stirred the most controversy, is perhaps the most different from previous versions of E.S.E.A., and is probably fraught with the greatest uncertainty: I refer, of course, to the new requirements concerning state academic standards and testing programs, the intersection of state testing and national assessment, the tracking of yearly progress, and the various interventions, rewards and sanctions that are wrapped into what's generally called "accountability" at the state, district and school levels.

No matter what one thought of the President's initial proposal (which I happened to like a great deal) or of the compromises and alterations that Congress worked in it (many of which I didn't like nearly so much), NCLB is now the law and I expect that everyone wants it to work effectively in carrying out its stated purposes, which including boosting student achievement, improving schools, giving people better information and closing some long-lasting and troubling gaps, so that, indeed, no child will be left behind.

The standards, testing and accountability provisions are at the core of this hope and this promise. But they turn out to be complicated. And somewhat mysterious. We actually don't know quite what is going to happen in the implementation of this law. That's partly because Congress left many important decisions to the Education Department and to the states and we don't know how they're going to handle these; partly because we're worried by the cautionary tales of weak and uneven implementation of past rounds of E.S.E.A.; partly because this is a country in which people hold different ideas of what constitutes good education and what's reasonable to expect from schools; and partly because NCLB embodies an idiosyncratic set of compromises between what the fifty states have discretion to do differently and what must be done uniformly by all of them. To recall just the most obvious of many examples: under NCLB, states are free to set their academic proficiency bars wherever they like but, whether they set them high or low, and no matter where their students are today in relation to those bars, they all have the same twelve years to get all their children over those bars.

These papers begin to explore such mysteries in the upcoming implementation of NCLB. We at the Fordham Foundation began the project with the premise that everyone wants it to work but that there's no unanimity on how that can or should happen and plenty of reason to worry about things that could go wrong, come unstuck, not be done at all, be done badly, not be foreseen, etc.

So we asked seven smart people (two of whom have co-authors, making for eleven smart authors) to examine some of these issues. We asked that their papers be written fast and kept short and accessible to ordinary readers. Most of the authors did most of those things. In fact, on the whole, they did a pretty terrific job.

Which doesn't mean they necessarily agree. There are interesting differences of view just among these seven papers. For example, Lisa Keegan and her colleagues are more bullish about what can be done with norm-referenced tests than Matt Gandal is.

We don't necessarily agree with them, either. I would come down differently on some issues. And some, indeed, are so intricate that another smart author, looking at the same issue, might have a different view of what the law provides and what the available data show.

Some of these differences came out in the lively discussion at the February 13 conference where, along with the authors, we were joined by five very able commenters, by the equally able Undersecretary of Education, Gene Hickock, and by 140 savvy and engaged education policy watchers, participants, analysts and journalists. It was lively, probing and sometimes a bit contentious. We wish more people could have been there.

The seven papers themselves provide much of the grist for that discussion, however, and they are worth the time and attention of anyone interested in the implementation of NCLB. It's a moving target, of course, The Education Department is already gearing up for "negotiated rulemaking." Much is in flux. That's why we concluded that getting these papers—some of them still working drafts—into cyberspace as quickly as possible would be more helpful than slowly trundling forth with a fully edited report of the traditional sort.

Reader comments and feedback are cordially invited. Let me emphasize that we're putting forth no "position" with these papers and have no political agenda. This is part of an earnest effort to begin reading the entrails of the No Child Left Behind Act in the hope that, if we understand them better, and are smart about what can and should and shouldn't happen, maybe we can boost the odds that this will indeed work well for American children, especially the neediest among them.

Chester E. Finn, Jr., President
Thomas B. Fordham Foundation
Washington, DC
February 2002

Multiple Choices: How Will States Fill in the Blanks in their Testing Systems?

Matthew Gandal

If someone had told me a couple of years ago that, over the next few years, every state was going to institute a grade-by-grade testing system, I would have laughed and thought that person was out of touch with reality and, frankly, politically naïve. Most states hadn't even established academic standards in each grade, let alone tests, and some were experiencing significant resistance from educators in the few grades where they were already testing. In a good number of states, moreover, policymakers did not believe grade-by-grade testing was necessary or desirable. Why would they all move to an annual testing system and how in the world would they pull it off?

What I hadn't considered was the confluence of events that would lead to the reauthorization of the Elementary and Secondary Education Act: a Republican president who believes in testing and accountability from a state that has shown that grade-by-grade testing can help raise achievement; his ability to get key members of his own party in Congress to stop viewing state standards and tests as an intrusion in local control of schools but rather a lever to improve them; and the leadership of key Congressional Democrats, who have come to see the power of standards and tests as a tool for achieving greater equity in American education and improving the life chances of the poorest children.

Now that the legislation has passed and the bill has been signed by the president, the question remains: how are states going to pull it off? The new ESEA amendments require states to begin administering annual tests in grades 3 through 8 in reading and math by the 2005-2006 school year. The previous law required states to test all students in those subjects but only twice within that 6-grade span. Only 16 states currently have grade-by-grade tests in reading and math, and only 9 of those states have tests aligned with their standards (a requirement of the law). The rest will have to fill in the blanks with new tests. Achieve estimates that well over 200 new state-level tests will have to be created over the next several years to meet the new federal requirements.

3 Big Questions

States have made great progress over the last ten years in setting academic standards for students and communicating those expectations to schools and parents. Most states have also tried to align their assessment systems with their standards so that what they are testing becomes more transparent for educators and parents and so that whatever "stakes" are attached to the test results are matched by reasonable opportunities for children actually to learn that which they're being held responsible for knowing. There is still considerable room for improvement, to be sure. But the groundwork is in place in nearly every state. As states move forward to fill in the gaps in their annual testing system, it is critical that the quality of the new tests and their alignment with state standards not get sacrificed.

Are States Ready?

Are states ready to respond to this challenge? It's too soon to be sure. Some states already have tests in all but one or two grade levels, so they only have to create a few new tests. But most states will have to more than double the number of tests they are now giving, and in doing so they will face both educational and political challenges (and incur financial costs as well). The educational challenges have to do with the quality of the tests and their usefulness in improving teaching and learning. This is something that states are already struggling with. The political challenges involve state and local control tensions and sustaining support from educators, parents, and business and community leaders.

Optimally, states will view the federal legislation as an opportunity to take a fresh look at their standards, assessments, and accountability systems and do what it takes to strengthen them. The goal should not simply be to fill in the blank years with tests so that every student is being tested in every grade. Rather, the goal should be to intelligently craft an assessment system that provides teachers, schools, and parents with the data they need to focus attention and resources and achieve better results.

Is the Market Ready?

Directly related to the question of state capacity is the capacity of the testing industry. One of education's dirty little secrets (made less secret last spring by a series of investigative reports by *The New York Times*) is that four major publishing companies have a virtual monopoly on the state testing market. While a few smaller firms have made some inroads over the last several years, the "big four" dominate this $700 million a year industry, creating and administering the tests in most states.

This raises some urgent questions: do these few companies have the capacity to develop over 200 new tests in a very short period of time? The normal cycle for creating a new assessment *in just one state* is 2-3 years. This now needs to happen in two subject areas in multiple grade levels *in at least 34 states*! In order to meet this demand, will the companies be forced to sacrifice their own (variable) standards of quality? Will they end up recycling old test questions and putting together hasty processes for creating new questions, thereby lowering the quality and sophistication of the assessments?

Is the Public Ready?

No matter how states approach the development of their new assessments, their greatest challenge by far will be sustaining the support of educators, parents, and the broader public as the new tests and accountability measures get rolled out. In poll after poll, parents, voters, taxpayers and opinion leaders have said they support testing, even high-stakes testing, because it provides them with some assurance that schools are effectively teaching and students are successfully learning. Educators have been less staunch in their support. They generally agree with raising academic standards, and acknowledge that tests are needed to measure achievement, but their support has begun to waver as real accountability measures have been put in place.

State and local policymakers will need to be mindful of this as they contemplate how to fill in the gaps in their testing programs. Few educators relish the idea of adding more tests on top of those they already have. States will need to be strategic: as new state tests get added, duplicative local tests should be taken away. And educators are sure to pay attention to what the new tests are measuring. The narrower and less sophisticated the questions, the more we will hear complaints from teachers that they are being forced to water down—or narrow—their teaching and focus on a test-prep curriculum.

The Challenge Ahead

At its core, the new law challenges states to measure student achievement more often in order to ensure that students are progressing on a path to proficiency. The idea is not to wait several years before taking the students' academic temperature, but rather to do it in every grade. More frequent testing leads to more frequent feedback to teachers, students and parents. And that feedback should allow schools to focus instruction where it is most needed and address achievement gaps for the benefit of all students. It is also intended to enable policy makers to intervene in situations where the testing reveals inadequate progress being made.

There are, however, a number of challenges to making this work as conceived, and although the law lists some important criteria state assessments will need to meet, Congress has left many of the toughest decisions to the U.S Department of Education and to the states themselves.

As states fill in the gaps in their testing systems, here are some of the things to watch out for: Will the new tests be adequately aligned to state standards? How challenging are those standards--are they worth aligning to? Will the new tests be aligned with existing tests, such that they measure a logical progression of skills from 3rd to 4th grade, from 4th to 5th and so on through 8th grade? Will the tests be sufficiently challenging? Will they measure advanced concepts as well as basic skills? Will the results be comparable across school districts within each state? How rigorous an approach will each state take to defining what it means to be "proficient"? How quickly and effectively will states report scores back to schools and households? Will states be mindful of the testing burden and work with districts to ensure that, as new tests get created, old ones head for retirement?

The governors, business, and education leaders who attended the 2001 National Education Summit last fall anticipated many of these issues and committed themselves to a set of principles that, if followed, will lead to stronger assessment and accountability systems. States that successfully address these challenges will end up taking maximum advantage of the opportunities the new law affords. Those that do not may very well end up taking a step backward in their reforms.

Testing Principles adopted at 2001 Summit:

- *Quality* – State tests should be designed to measure student progress against clear and rigorous standards. Reports sent to schools and parents should indicate how students perform against the standards — not just how they compare with other students. Tests developed for other purposes cannot meet this need. The tests should measure the full range of knowledge and skills called for by the standards, from basic to most advanced.

- *Transparency* – In a standards-based system there should be no mystery about what is on the test. Students, parents, and teachers should know what is being tested. They should be confident that if students are taught a curriculum that is aligned with state standards, they will do well on state tests. The best way for states to ensure transparency is to publicly release questions from previous years' tests, along with sample student answers at each performance level.

- *Utility* – Ultimately, it is the clarity of the results and the manner in which they are used that will make a difference in schools. Test results should be returned to schools and parents as quickly as possible without compromising the quality of the test instrument. Score reports should be clear, jargon-free, and designed to guide action.

- *Comparability* – The goal of state assessment programs is to create measurement systems that can accurately track and compare student and school progress from year to year. To accomplish this, the tests from one grade level to another must be aligned with state standards, and the results must be comparable from grade to grade so that student progress can be tracked from year to year.

- *Coherence* – State tests are only one piece of a comprehensive data system. Local and teacher-developed assessments are important too. States must work with districts to ensure that all tests serve a distinct purpose, redundant tests are dropped, and the combined burden of state and local tests remains reasonable.

- *Strategic Use of Data* – Closing the achievement gap can only occur if student achievement data is disaggregated by race and income, and if schools are required to show that all groups of students have made reasonable progress. By regularly reporting how every school is performing against state standards, states can focus attention on the problem, on the progress that some communities and schools are making in response, and on areas where additional work is needed.

How Will States Respond? Four Scenarios

While ESEA lays down some clear markers on issues of academic standards, testing, and accountability, states have numerous options in determining how to fulfill the requirements. The Department of Education will either need to get much more concrete about what is expected or the states will end up determining the answers to these questions themselves. It is worth playing out several plausible scenarios to highlight the costs and benefits of the different approaches states might take.

Scenario #1—Cheap and Easy

It is more costly and time consuming to create new tests aligned with state standards than to take existing tests off a publisher's shelf and assert that they are aligned. The fastest, cheapest way for states to fill in the gaps in their testing programs is to purchase ready-made tests such as the Stanford 9, Iowa Test of Basic Skills, and Terra Nova. These are in widespread use in schools today, but they are not designed to measure student attainment of any particular state's standards. Rather, their main purpose is to compare one student's achievement against that of other students in a national sample, in essence comparing that child against an average.

Comparing pupil performance to an average or "norm" is very different than measuring whether or not that child has met a specific set of academic targets. The targets, or standards, provide something for students and teachers to aim for, and those standards do not fluctuate based on how other children are doing.

Although it is not impossible for commercial tests to be well aligned with states' standards, it is highly unlikely. In studies that Achieve has conducted for states, we have found that commercial tests typically touch on some standards but miss the mark on others. The pattern is that commercial tests tend to focus on what is easiest to assess, and it is often the most rigorous knowledge and skills that are not adequately measured. The result is a testing system that is out of sync with what states profess they want students to learn.

If, therefore, states opt to use "off-the-shelf" tests to fill in the grades where they do not currently have tests, they will likely sacrifice the measurement of their standards in those grades. A combination of customized tests in some grades and off-the-shelf tests in others may also end up sending mixed signals to schools and parents about what students are expected to learn. If, for example, a state uses customized tests in 4th and 8th grades and off-the-shelf tests in the other grades, the 4th and 8th grade teachers may end up paying attention to the state standards because that is what is being tested, but the teachers in the other grades may pay less attention to the standards and more attention to what's on the commercial tests. Imagine a school trying to organize its curriculum in such an environment; imagine teachers trying to collaborate across the grades; imagine parents trying to make sense of their children's test scores from grade to grade.

There is a twist on this strategy that a few states have pursued. In order to get a testing system in place quickly, California began in 1998 by adopting a series of off-the-shelf tests for grades 2-11 (the Stanford 9) and then worked with the testing company (Harcourt Educational Measurement) to adapt or "augment" those tests over time to align better with the state's own standards. Starting in 1999, California children began taking the augmented version of the tests, called "STAR" exams (Standardized Testing and Reporting System). These exams consist of a combination of questions from the Stanford 9 and new test questions that were added to reflect the California standards. According to state officials, as many as 75% of the test questions in math had to be created from scratch to align with the standards; a smaller number of new questions were needed in English.

Although education officials in California readily admit that their unorthodox approach caused confusion and even skepticism in schools across the state, they seem optimistic that their transitional strategy will result in tests aligned with their standards. Before other states consider trying this approach, though, it is worth a more careful look: Just how different are the "augmented" tests from the original ones? How well do they in fact align with the state standards (which, by the way, are among the most rigorous in the nation)? If they do, in fact, align well, how much of that has to do with the fact that California's size and market share allowed it to push the testing company harder than a typical state could? Most states find that they have little leverage over these companies, but big states have greater influence due to the size of their student populations and the huge markets that get opened up for textbooks and other products.

The truth is, alignment of tests with standards is difficult to achieve. Even states that have created their own tests from scratch have had a hard time measuring their standards well. But getting it right will be essential if the new assessments that states create are to add value to the existing ones, and become tools that teachers, parents, and policymakers can rely on to raise student achievement. Doing that well is not apt to be cheap.

Scenario #2—Leave it to Districts

As state leaders have pondered how they're going to fill in the grades where they currently do not have tests, some have said that they would rather let districts use their own local tests in the years when the state does not test. This is clearly the most politically convenient solution, as it sidesteps the state/local tensions and allows districts that already test students in grades 3-8 to leave those tests in place. It does, however, raise serious questions about the comparability of data across those districts.

Formal studies by the National Research Council and informal studies by Achieve have concluded that it is nearly impossible to compare results of different tests in any meaningful way. This is because different tests measure different concepts and skills, so proficiency on one test rarely translates to proficiency on another. If states were to pursue this path of least resistance, therefore, they will likely sacrifice the ability to compare achievement results across districts in the grade levels where the state itself does not test. How important is this to states? Will the lack of a common test in each grade skew the accountability system? Which tests will be factored into the adequate yearly progress formula: the state tests, the local tests or both? How can one provide cumulative results for the state as a whole if the tests differ from place to place within it? Wouldn't that lead to data that are very difficult to disaggregate? Will multiple tests send conflicting signals to schools as to where they should focus their curriculum and instruction?

Scenario #3—New Customized Tests

In order to stay true to the principles of alignment, coherence, and comparability, the most desirable strategy for building an annual testing system is for states to develop new tests for the grades where they don't have them. Those tests would be both aligned to the their academic standards and aligned with the tests that they already have.

There are several different ways states might approach this. Some may choose to match the length and sophistication of their existing tests. Other states may decide to alter the format and length of their new tests. They may do this to reduce costs, to reduce the amount of time needed for students to take the tests, or to make the tests more diagnostic and useful to local educators. This is where a creative approach to the task could have the greatest educational payoff.

Imagine a state that currently has reading and math tests in 3rd, 5th, and 8th grades, and each of those tests is 90 minutes long and consists of a combination of multiple-choice and extended response questions (i.e., questions requiring written answers, such as essays). Confident in the data those existing tests provide and wary of the costs of producing identical tests in new grades, state officials might decide to create a shorter version for grades four and seven designed to provide a brief snapshot in between the other tests. The new tests might have fewer questions or rely more heavily on multiple-choice questions, and might only require 45 minutes of test-taking time. This approach would allow states with sophisticated assessments to maintain them at some grades while using more economical versions at other grades.

Another approach might be to make the new tests as sophisticated as the existing tests, but to get creative in how they are scored. Indiana is one state considering this. The idea officials are exploring is to have classroom teachers scoring certain portions of their students' tests and to make the results immediately accessible to schools and parents. There would clearly be quality control and consistency issues that the state would need to work out, but in addition to saving money on centralized scoring, one of the benefits of this approach is that teachers would be much more invested in the assessment process and, therefore, may end up using the results in their classrooms. In fact, done right, grading state assessments could be a very effective form of professional development. Indiana is also exploring the development of formative assessments that teachers can voluntarily use at any point during the school year to determine how their students are advancing toward the state standards.

However states approach the task of creating new tests, it is critical that they remain vigilant about test quality. Achieve's work has revealed that even states that have created their own assessments for the purpose of measuring their own standards have had a difficult time getting it right.

Scenario #4—State Collaboration

When it comes to creating high quality tests worth teaching to and basing serious accountability systems on, the deck is clearly stacked against most states. High quality tests cost more to create and there is a limited pool of talent available to help them accomplish this. Given these tensions and the real pressure that states are under to get so many new tests in place relatively quickly, it is legitimate to ask why states need to go it alone.

The most logical strategy for responding to the ESEA testing requirements is for states to pool resources and develop common assessments that they can share. This would allow states that do not have the market power of California, New York, and Texas to work together to leverage better quality tests. They are all relying on the same few companies to create these tests. Why not step back, form strategic partnerships, and leverage the situation?

There are three reasons that states should consider doing this. The upsides are better quality tests, lower costs, and more comparable data across states since they will be using the same tests. The cost savings could be significant at a time when state budgets are tight and it's not clear whether Washington is earmarking enough money to offset state testing costs. The comparability advantage also deserves more attention than it typically gets: one reason the legislation requires all states to give NAEP reading and math assessments every two years is that policymakers want better ways to compare results across states against a common standard. Why not build that comparability into states' own assessment systems while they have the chance? This happens to be the reason some state policymakers and parents like the idea of using norm-referenced tests—it gives them some ability to compare results beyond their state.

The new law specifically allows states to form consortia and pool resources to create and use common tests. The main thing standing in the way at this point seems to be habit. States are used to working individually with test publishers to create their own tests. They are not used to a collaborative approach. This may change as states look ahead at the need to build over 200 new tests.

There is at least one consortium already in place that could be very helpful to states as they develop their ESEA strategies. At the request of governors and education commissioners in a number of states, Achieve launched an initiative in 1999 known as the Mathematics Achievement Partnership to help states work together to raise mathematics standards and achievement. Fourteen states are currently involved in the partnership, which will provide them with an internationally benchmarked 8th grade math assessment, tools for improving the middle school math curriculum, and strategies for improving the professional development of middle school math teachers. We are exploring how states can tap into the consortium to develop tests in the grades where they currently do not have them.

Getting It Right

The task ahead for states in building an annual testing system reminds me of what must be a fairly typical challenge facing city planners when they address changes in traffic patterns. Oftentimes, heavier usage on some roads necessitates adding stop lights at more intersections to control traffic and ensure safety. When confronted with the challenge of adding traffic lights at more intersections along a busy street, what would a thoughtful city planner do? Would he purchase the least expensive product even if the signals it sent were different than those of the existing traffic lights? Would he ask the residents on each block to build or buy their own traffic light? How would traffic be affected if the new signals were not timed with the existing ones? Would it help control the flow of vehicles or simply confuse and frustrate drivers and pedestrians?

The thoughtful city planner keeps the endgame in mind as he devises his plan. The goals are safety and the smooth flow of traffic, not placing a traffic light at each intersection. That's simply a means to the end. If poor decisions are made, it is quite possible that the addition of lights at each corner could make the streets more congested and less safe.

It is the same with building an annual testing system. Approached intelligently, grade-by-grade testing can be a real improvement over what many states currently have in place. But not all strategies for creating annual tests will result in a coherent assessment system. States must take care to get it right.

The President and Congress did make an effort to address some of the issues discussed in this paper. There are a series of criteria laid out in the law that state assessment systems will need to meet. These include: alignment with state standards; reporting scores for each individual student; disaggregating the data by race, ethnicity, and socio-economic status; providing itemized analyses pointing to students' strengths and weaknesses in each particular skill area; returning the results before the beginning for the next school year; and assessing "higher order thinking skills and understanding."

At this stage, the question on most people's minds is how rigorous federal officials will be in their interpretation of these criteria and, more importantly, how serious they will be about enforcing them. Federal officials can and should play an important role in clarifying criteria and reviewing state plans, and if they take a hard line on some of these important issues, states could be left with a smaller but smarter set of options.

If past experience is our guide, however, we should not expect the federal government to fully solve complex issues such as the quality, alignment, comparability, coherence and utility of state standards and assessment systems. The federal government can lay down clearer markers and use the bully pulpit, but in the end, these are issues that state leaders must address for themselves.

Using NAEP to Confirm State Test Results: An Analysis of Issues

Mark D. Reckase

The new Elementary and Secondary Education Act (ESEA) amendments, "No Child Left Behind," require that the National Assessment of Educational Progress (NAEP) reading and mathematics tests be administered every other year in grades 4 and 8. Further, states must participate in the component of NAEP that is used to obtain estimates of students' academic performance at the state level. This part of the NAEP program is called State-NAEP. Participation in State-NAEP has been voluntary in the past, but the ESEA amendments make participation a condition of accepting Federal funds related to the legislation. While the legislation does not indicate what is to be done with the results of NAEP testing, it does imply that NAEP will be used as a check on the reading and mathematics assessment results reported by each state. Further, states will be required to administer their own reading and mathematics assessments to their students every year in grades 3 through 8. The purpose of this policy memo is to summarize the issues related to the use of NAEP to confirm the assessment results reported by states.

Testing Programs in the ESEA Legislation

A Brief Description of NAEP

NAEP is an extensive program of data collection that includes achievement tests in a number of subjects, including, but not limited to mathematics and reading. NAEP also collects information about characteristics of the student population and features of the educational system. NAEP results, and the many interpretive reports produced from those results, provide an ongoing description of the functioning of the educational systems in the United States.[1]

NAEP tests are uniquely different from state assessments in a number of ways. First, the tests attempt to measure student capabilities (what students know and can do) on a domain of process and content knowledge that is common to the state educational systems across the United States. The creators of the document describing what is included in that domain also attempt to include content and processes recommended in future-oriented standards documents (e.g., those promulgated by the National Council of Teachers of Mathematics) so that the domain definition will be applicable for a number of years into the future. Allowing the national standards documents to influence the domain definitions implies that states are expected to move their curriculum in the direction of those standards.

---

[1] Details of features of NAEP are presented in a number of documents including Braswell, Lutkus, Grigg, Santapau, Tay-Lim and Johnson (2001).

The domain of coverage for a NAEP subject matter area is described in a document called a "framework" (e.g., *Reading Framework for the 1992 National Assessment of Educational Progress* (NAGB (1992)). A consequence of the need for NAEP to be appropriate for assessing student performance in all states is that it can not focus too closely on the educational goals from any one state. NAEP assesses the common core of all state programs, but it does not assess the instructional goals that are unique to individual states.

A second way that NAEP is unique is that no student takes the entire test. Because NAEP endeavors to assess what students know and can do in a very broad domain, the full NAEP tests contains a large number of questions --145 to 160 questions for NAEP Mathematics, for instance. This number of questions is too large for any student to attempt in a reasonable period of time. To keep thorough domain coverage, but also keep the testing time to a reasonable amount, each student takes only 36 to 45 mathematics questions. Test booklets contain overlapping sets of questions so that the results from all of the examinees can be combined to determine the expected distribution of performance on the full set of questions for the full sample of students. However, it is not possible to obtain a good estimate of performance on the full domain of knowledge and skills for any individual student because the student has responded to only a small part of the entire test.

A third unique feature of NAEP is a direct result of the item and student sampling approach that it uses to keep testing demands within reasonable bounds. Because students take only part of the test, no student scores are reported. Also, tests are only administered to a random sample of students from the nation and from within participating states. A consequence of the sampling approach is that only estimated score distributions for state and national groups can be reported. NAEP summarizes the information from these distributions using percentages above achievement levels set by the National Assessment Governing Board (NAGB) and descriptive statistics (means and standard deviations). It is not possible to track individual student's performance on NAEP over years or directly compare student performance on NAEP with that student's performance on a state test. Nor is it possible to report NAEP results at the school building level because only a small number of students from any school take the test, and those students take only part of the full set of test questions.

The unique features of NAEP have not interfered with its use as a general indicator of the quality of education in the United States. However, they will need to be taken into account when NAEP results are compared to state results.

State Assessments

State assessment procedures are notable for the diversity of approaches that they take. Some states purchase existing tests from commercial test publishers as all or part of the state assessment program. This approach would seem to indicate that these state education officials believe that the commercial tests are sufficiently aligned with the curriculum and instruction goals for the state. Other states hire test development contractors to custom develop elaborate assessment programs according to state developed test specifications. The test specifications for these programs vary greatly. Some include performance assessment tasks that are scored by commercial companies, others are multiple-choice only, and some use computerized testing procedures as part of the assessment program. One state (Iowa) does not have a state assessment program, though most students in the state take the Iowa Tests of Basic Skills and Iowa Tests of Educational Development at some point in their schooling.[2]

The diversity of state assessment programs provides a challenge for the use of NAEP to confirm the results of those assessments. The state assessment programs have different content, schedules for administration, purposes, stakes, and technical characteristics. Further, many of these features will likely change in response to the ESEA legislation. At the very least, many states will have to increase the frequency of testing in grades three through eight in reading and mathematics. The next section of this memo highlights a number of the more important issues related to the use of NAEP for confirmation purposes. The following sections discuss the effects of differences in state testing programs on the interpretation of NAEP/state assessment comparisons.

The Relationship between NAEP and a State Assessment

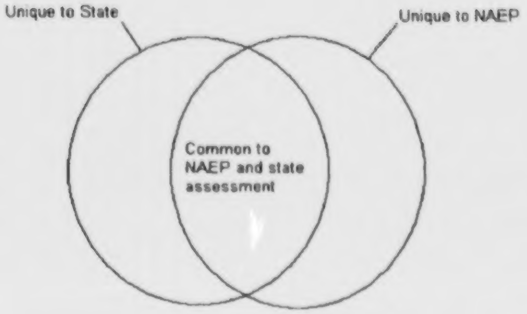
Domain Overlap

The starting point in the design of an achievement test is the specification of the domain of content and skills to be covered by the test. In theory, there should be a description of the domain at a level of detail that will allow an interested party to determine whether a specific test task should be on the test because it measures part of the domain, or whether it should be excluded because it does not. The NAEP framework documents are good examples of domain specifications. Unfortunately, like everything else with state assessment programs, descriptions of domains vary substantially across states. Some give very general statements of academic goals; others provide detailed descriptions of desired academic content and skills.

A key to determining the comparability of NAEP and state assessment results is an evaluation of the commonality of the target domains. The following diagram gives a simplified representation of the overlap in those domains. The content domain for a state assessment program is represented by one circle and the domain for NAEP is represented by another circle. Within a circle is the content and skills to be measured by an assessment program. Outside the circles are the content and skills that are not included in the domains for either of the two assessment programs. For each assessment, there is part of the domain that is in common with the other assessment and part that is not.

---

[2] A brief summary of state assessment programs is given in Meyer, Orlofsky, Skinner and Spicer (2002).



States vary in the amount that their assessment domains overlap with NAEP. For some, there is almost complete overlap. For others, the overlap is modest. Unfortunately, there do not seem to be any formal studies of the amount of overlap between domains for NAEP and state assessments. Such studies would be major undertakings that would require in-depth analysis of every state testing program. There would be a further complication that state assessment programs are not static – they change frequently, sometimes because of changes to the curriculum, but also because of other factors such as the need to reduce costs, or because of changes in educational policy within the state.

Assuming that the amount and composition of domain overlap can be determined, a critical issue when comparing NAEP and state assessment results is the part of the domain that is emphasized by instruction within the state. Although a state may define a large content and skill domain as the focus of instruction, not all parts of that domain will be treated with the same emphasis in every classroom. If the focus of classroom instruction is on parts of the state's domain that do not overlap with the NAEP domain, then student performance may improve and be documented on the state assessment while that improvement is not shown on NAEP. NAEP might even show a decline if the part of the domain that is common to the assessment programs and the part that is unique to NAEP are given little instructional emphasis.

To the extent that NAEP has captured the important outcomes of the nation's educational systems, the cases of low domain overlap and of instructional focus on things not covered by NAEP should be rare. But it is possible that a state could show improvement, NAEP could show decline, and they could both be correct because instruction is focusing on different parts of the combined domain for the two tests.

Performance Standards

NAEP reports results in two ways. The first is estimated test score distributions on the NAEP standard score scale. This type of reporting includes mean scores for demographic groups and state samples. The second way that NAEP results are reported is percentages above achievement levels set by NAEP's governing body, the National Assessment Governing Board (NAGB). NAGB has set three such levels labeled basic, proficient, and advanced. The achievement levels are ranges between cut scores on the NAEP score scale. NAGB considers these cut scores as definitions of performance goals for what students should know and be able to do at grades 4, 8, and 12.[3] The NAGB achievement levels take on special meaning in the ESEA legislation because the legislation specifies that states must define their own "proficient" and "advanced" levels, as well as a "basic" level. The language of the legislation uses the same labels already used for the NAGB-developed achievement levels on NAEP.

States also set cut scores on their assessments, but even when they use the same labels as the NAGB achievement levels the meanings of the state standards might be quite different. For example, a state may use the term "proficient," but in terms of the number of students who attain that level or higher, the state's proficient level may be similar to the NAGB "basic" level. Such differences in meaning of state and NAGB standards are not likely a sign of duplicity. The research on standard setting shows that different standard setting methods, different statements of policy, and standard setting panels with different characteristics are likely to produce different standards.[4]

The location of cut scores on a score scale is important because the location indicates where the reporting system will be sensitive to changes in student performance. Consider the following thought experiment. Suppose that a standard is set on a mathematics test by placing a cut score for reporting at roughly the level of difficulty of simple addition problems. Also suppose that at grade 4 in one school, the students are not yet doing well on addition, while at another school most of the students have mastered addition. In the first school, if instruction focuses on simple addition, many students will move from below the standard to above the standard. It is likely that the percent above the standard will improve quite dramatically. In the second school, however, because the students already know the material and because instruction is focused on other, probably higher level skills and knowledge (e.g., fractions), the increase in percent of students attaining that state's standard in that school will be small. The opposite effect can occur if the cut score is set at a level that is consistent with the difficulty of the fraction problems. In that case, the second school would show a lot of improvement and the first school would show very little.

The NAGB "proficient" level is a fairly high standard. Changes in the percent above that standard will likely reflect achievement gains for students whose instruction focuses on the more difficult NAEP content. Changes in the proportion above "basic" will likely show improvements for students whose instruction focuses on relatively easy NAEP content.

---

[3] For a discussion of the issues related to the standards set by NAGB, see Bourque and Byrd (2000).

[4] See Cizek (2001) for recent information on standard setting.

Context of the Assessment

Not only do NAEP and state assessments differ on domain coverage and the placement of performance standards, they also differ in the context for the assessment; that is, the way that the assessment is perceived by the students and the local school district staff. For example, some states use their assessments to determine whether students will be promoted to the next grade or whether school staff will receive monetary awards for helping students reach instructional goals. These assessment programs are called "high stakes" because there is a direct and important consequence to the students and school staff. In such cases, it is likely that students will be motivated to do well and the school staff will do what they can to help the students perform at their best.

The amount of "stakes" for state assessments varies quite dramatically. Some states use the assessment results only for general school accountability purposes with no direct consequences for students. Some states test a sampling of students rather than every student. Other states make the assessments a very important part of the state instructional system. Teacher salaries may depend on the assessment results and students may receive direct rewards or punishments. The high level of variability across states with regard to "stakes" adds to the complexity of comparing state results with NAEP results.

NAEP has no direct consequences for students or school staff because NAEP results are not reported at the school or student level. Students do not receive scores and schools do not receive summaries of student performance. These features of NAEP make it a "low stakes" assessment at the school and student level. The differences between contexts for state assessments and NAEP need to be taken into account when interpreting comparative results.

Analysis

When comparing state assessment results with NAEP results for a single curriculum area, there are nine possible results as depicted in the cells in the following table. NAEP confirming state results would seem to require that both testing programs have results in the cells with the Xs. The question of concern here is "How likely is it that NAEP and state assessments will give results in these cells?" To answer this question, all of the issues that have been summarized need to be considered.

		State Assessment		
		Decline	Stay Level	Increase
NAEP	Decline	X		
	Stay Level		X	
	Increase			X

First, the issue of domain overlap needs to be considered. For most states, the domain overlap between NAEP and the state assessment will be at least moderate. NAEP was designed to measure the common content of the instructional systems of all of the states. Unless a state has instructional goals that are notably different than those of other states, there should be some commonality between domains of coverage for NAEP and a state assessment. However, it is not likely that the overlap will be total for any state. It is possible that there may be important parts of a state domain that are unique to the state and not included in the content of NAEP. If the state focuses instruction and assessment on the unique features to the exclusion of the common components, it is possible for the state assessment to show gains when NAEP does not. It is also possible for NAEP to show gains when a state assessment does not if instruction focuses on the unique features of NAEP (e.g., instruction may be focused on national curriculum standards) rather than the unique features of the state assessment. This seems less likely, but possible. The existence of these possibilities suggests that part of the interpretation of NAEP results for confirming state results will need to be a judgment of the overlap between the assessment domains. Substantial overlap makes NAEP a stronger tool for confirmation. Low overlap indicates that NAEP can not provide solid evidence for confirmation or disconfirmation.

Second, the context of the state assessment will also likely affect the usefulness of NAEP as a source of evidence for confirmation. If the state assessment is high stakes and NAEP is low stakes, students may try very hard on the state assessment and not very hard on the NAEP. Real situations may be more complicated. There are more possibilities than motivated and not motivated. Students vary in level of motivation and the level of student motivation may interact with the level of difficulty of items. Students may give a reasonable level of effort to easy items even when the test does not count for them, but they may give up on hard items when the test does not have direct consequences. The result of differences in stakes may be that students show improvement on the state assessment if it is high stakes and no improvement or a decline on NAEP.

The context of state assessments and NAEP may differ in other ways that may affect the comparison of results. The assessment programs may be administered at different times of the year. If the state assessment is administered in the fall, and NAEP is administered in the spring, the amount of exposure to the curriculum will differ. The differences in instructional time will influence the amount that students have learned by the time the test is administered and the amount of gain that can be detected. The quality of the assessments may also differ, affecting the confidence that can be placed in the reported results.

The location of standards on the assessment can result in similar differences in results. Students at all points in a distribution of performance will not likely improve by equal amounts. If a school focuses on the improvement of basic skills, performance standards set at a relatively low level will show the greatest change in the percent attaining those standards. The NAGB "proficient" level is a high standard so it may not be sensitive to changes in basic skills. A basic skills oriented state standard might show improvement while the percent above NAGB "proficient" does not. The opposite may occur for schools focusing instruction at a higher level – NAEP may show changes when the state assessment does not.

A solution to this problem is to look at changes at all levels of student achievement rather than at single cut scores. NAGB is currently investigating reporting procedures for NAEP that can show changes along the entire NAEP score scale. These same procedures could be used by states as well.

The description of state and NAEP assessment programs given here is based on the current characteristics of those programs. However, the legislation will likely result in significant changes to both NAEP and state assessments. A recent review of state testing programs in *Education Week* indicates that only eight states currently meet the requirements set out in the legislation. Many states will have to expand their reading and mathematics assessments to meet the requirement of testing every year from grade 3 to grade 8. NAEP will also have to change its testing schedule to provide results every other year in mathematics and reading. While it is likely that significant changes in these assessment programs will occur, the full impact of the changes will not likely be understood for several years.

Conclusions

Jointly interpreting state assessment and NAEP results in a coherent way will not be a simple task. Many factors need to be taken into account when making such interpretations including the amount of content overlap, the location of cut scores on the score scales, and the context for the assessments. This is not to suggest that the joint interpretation of the test data is impossible or unwise. Experience from analysis of ACT and SAT college admissions tests and other testing programs indicates that tests constructed from different test specifications can yield highly correlated results. It is likely that NAEP results and state assessment results will be related as well. With careful consideration of threats to accurate interpretations and realistic judgments about the amount of effort that will be required to make accurate interpretations, joint use of NAEP and state assessment results should lead to better understandings of the functioning of the educational systems in the United States.

References

Bourque, M. L. & Byrd, S. (Eds.) (2000). *Student performance standards on the National assessment of educational progress: affirmation and improvements.* Washington, DC: National Assessment Governing Board.

Braswell, J. S., Lutkus, A. D., Grigg, W. S., Santapau, S. L., Tay-Lim, B. & Johnson, M. (2001). *The Nation's Report Card: Mathematics 2000 (NCES 2001-517).* Washington, DC: National Center for Educational Statistics.

Cizek, G. J. (Ed.) (2001). *Setting performance standards: concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Meyer, L., Orlofsky, G. F., Skinner, R. A. & Spicer, S. (2002). The state of the states. *Education Week, 21*(17), 68-169.

National Assessment Governing Board (1992). *Reading framework for the National Assessment of Educational Progress.* Washington, DC: Author.

Adequate Yearly Progress: Results, not Process

Lisa Graham Keegan, Billie J. Orr & Brian J. Jones

When President Bush signed the *No Child Left Behind Act of 2001* (NCLB) into law on January 8, 2002, he brought to the public school system a new demand. All students—regardless of race or socioeconomic status—must be held to the same academic expectations, and all students—regardless of race or socioeconomic status—must have their academic progress measured using a newly-refined concept of adequate yearly progress (AYP).[1]

The term AYP should be nothing new to educators. Title I of the previous version of the Elementary and Secondary Education Act, the *Improving America's Schools Act* (IASA) of 1994, introduced the concept of adequate progress in its requirements that all states establish academic content standards, develop tests to assess student progress in those standards, and create performance standards for those tests. But the focus of the 1994 law centered much more on the process of building the AYP mechanism that would be used to measure achievement in Title I schools and for Title I students than it did on ensuring actual academic progress for all students. Consequently, most states have dual accountability systems in place—one for Title I schools and another for all public schools. In 2000, only 22 states had a single, unified system to judge the performance of all public schools.[2]

With NCLB, all this changed. The play is no longer the thing; success in complying with the law will no longer be based upon whether a state has created academic standards and testing, but rather on how well all of its students are doing in making real progress toward meeting those standards. That means testing all students, and it means using the same system for all students; thus NCLB requires states to use a single accountability system for all public elementary and secondary schools to determine whether all students are making progress toward meeting state academic content standards.

This expectation defined by NCLB—that all children will make continuous progress toward proficiency on state standards—is the underlying motive behind the new AYP. The goal is to ensure that all students, regardless of what they look like or how much money their parents earn, make adequate yearly progress, period. "All students can learn" is no longer just a mantra, it's a goal that will be measured every year.

The AYP process sounds relatively straightforward: States set the bar for what is deemed "proficient" in relation to their academic standards. They must then define what level of improvement will be sufficient each year to determine not only whether districts and schools have made "adequate yearly progress" toward meeting the standard of proficiency, but also the rate at which they will get all students to proficiency in twelve years. Finally, after testing students each year, states will disaggregate the testing results to determine how specific populations of students are achieving at the state, district, and school levels, and make those results available to the public. This is simple in description, but complicated in execution—and, ultimately, central to the law. AYP is used throughout NCLB to determine compliance, rewards, and sanctions. Process is not enough; it's results that count.

---

[1] *No Child Left Behind Act*, P.L. 107-110, 107th Congress, 1st Session, 2001.

[2] Margaret E. Goertz and others, *"Assessment and Accountability Systems in the 50 States: 1999-2000"* (University of Pennsylvania: Consortium for Policy Research in Education, 2001), 30.

Precisely how we define results—even when it comes to such seemingly simple tasks as defining terms like *proficient* or *adequate*—will be decided in collaboration with the U.S. Department of Education and the states. While this law gives strong guidance, we would all do well to approach this collaborative process with humility. State accountability systems that seek to ensure the academic success of all students are still relatively new and unstudied phenomena. Our experience to date has given us much confidence that the broad infrastructure of NCLB is sound, but there is still much to learn and many ways to approach the requirements of this new law.

Defining a System: "Specific Ambiguity"

Under NCLB, Congress provided the states with significant flexibility in developing state accountability systems, and with greater flexibility in general program administration than has previously been permitted in federal education law. For example, State and local education agencies will be allowed for the first time to shift up to 50 percent of their non-Title I administrative funds between programs, or they may even shift these funds into Title I itself (though they cannot move funds out of Title I to other accounts). States can also apply to receive "flexibility authority," which will be awarded to seven states on a competitive basis to demonstrate even greater gains with greater freedom.

Consistent with this new flexibility, while the objectives of the AYP requirements in NCLB are obvious as general guidance, they leave a great deal of room for interpretation in their specific implementation. For this reason, the U.S. Department of Education will be issuing further instruction on many of the details of the law. We would advise those involved in the rulemaking and guidance process to proceed cautiously, for the very vagueness of the law—this "specific ambiguity"—is actually an asset, as it leaves each state room to experiment within its own strengths and limitations. Rulemakers should not eliminate the desired and intentional ambiguity of the law; rather, they should jointly be seeking ways to learn from it. As Thomas J. Kane noted in an analysis of the House and Senate AYP proposals,

> ...states are currently experimenting with a wide range of different types of accountability systems. They should be allowed to continue experimenting, until the Nation reaches a consensus regarding the ideal way to determine which schools are making

> adequate yearly progress and which are not.... [I]mpatience is an insufficient excuse for bad education policy.[3]

While NCLB defers in certain respects to state policies and practices, it does lay down some non-negotiable directives that states must adhere to in their efforts to develop an AYP process. One might compare this to a road map on which main thoroughfares and destination are clearly marked, but unmarked side streets and alleys are also open to travel along the way.

Under the law, each state is required to work with its teachers, parents, principals and local educational agencies to create a state plan that incorporates challenging academic content standards and student achievement standards that apply to all children within the state. The academic achievement standards (formerly called performance standards) must describe *basic*, *proficient* and *advanced* levels of achievement. As stated previously, this is crucial to understanding the concept of AYP, because the goal is for all children to reach the *proficient* level (or beyond). The state must also implement a single accountability system that ensures that its schools, districts and the state as a whole make adequate yearly progress.

Further, while each state is responsible for the specifics in defining how it will determine "progress," the federal law is clear that the state's definitions of AYP must have the same high standards of achievement for all public schools in the state, and they must follow a 12-year timeline for getting all students to proficiency. The state's criteria must be statistically valid and reliable, require continuous and substantial improvement for all students, and measure progress based on state reading and mathematics tests. Secondary schools must include graduation rates as a factor in determining progress, and elementary schools must use one additional indicator such as attendance, promotion rates or increases in participation in advanced classes.

Data from the 2001-2002 school year will establish the starting point for measuring the percentage of students meeting or exceeding the state's level of proficiency. States must set the initial bar at a level based on either its lowest achieving demographic group, or the scores of its lowest achieving schools, whichever is higher. However, regardless of where the initial bar is placed, states *must* define AYP so that *all* students in *all* groups are expected to improve and achieve the proficiency level in 12 years.[4] The law is specific in this goal, but ambiguous in the starting point, deferring to the states for the criteria they will use for the initial placement of the bar.

Once the starting level has been determined, states must then begin raising the bar over time, increasing the number of students meeting or exceeding the state's level of proficiency over time, with the goal being 100% of students at proficiency in 12 years. The statute requires that the bar be raised in equal increments over time, and must be raised for the first time not later that two years into the process, and then again at least once every three years. Where states have leeway is in determining the initial "height" of the bar, and the rate at which it will be raised over time until 100% of students reach proficiency.

---

[3] Thomas J. Kane and others, "*Assessing the Definition of 'Adequate Yearly Progress' in the House and Senate Education Bills.*" (Los Angeles: School of Public Policy and Social Research, UCLA, 2001), 12.

[4] *No Child Left Behind Act*, P.L. 107-110, Section 1111 (b)(2), 107th Congress, 1st Session, 2001.

Finally, to ensure that the most disadvantaged students do not get left behind in this process—so that states and schools don't get the more affluent children to proficiency first, then go back and start working on at-risk children in the waning years of the 12 year deadline—states must include separate measurable objectives for "continuous and substantial improvement" in both reading and math for students who are minorities, poor, disabled, or of limited-English proficiency (LEP). This is how states can monitor how well they are doing in closing the achievement gap.

The bottom line is that, in order to demonstrate adequate yearly progress, the state and its districts must show that schools are meeting or exceeding the state annual measurable objectives for all students and for students within each subgroup.

It is important to note that there is also a "safe-harbor" provision found within NCLB, created to address the concern that too many schools would be identified as failing simply because one subgroup—for example, LEP students—failed to meet the state AYP goals. This provision allows schools to avoid being considered as failing so long as (in this particular example) the number of LEP students who are below proficiency decreases by 10 percent when compared with the proceeding year, and if LEP students also made progress on one or more of the additional academic indicators listed above. The law also requires at least 95% of students enrolled in the school and in each subgroup take the state tests in order to meet the standards of AYP.[5]

As an external audit for states to gauge the quality of their own standards—to give them some idea of how high their bar for proficiency is set and how well they have defined progress toward that bar—states will be required every other year to administer the National Assessment of Educational Progress (NAEP) tests in reading and math. This is not only a significant change from prior law (where NAEP was optional and administered only once every four years) but a critical one. NAEP results will act as both light and leverage for states serious about taking a closer look at their standards and making any necessary modifications to ensure that they remain rigorous.

What will an ideal system look like? Frankly, we're not sure yet. Clearly, states will develop a single accountability system for all students, create definitions of progress that fall within federal parameters, and lay out a timeline for getting all students to proficiency in 12 years—and there end the details. Through NCLB, the federal government has said, "Here are the guidelines, the flexibility, the resources, and the expectations. We'll meet you back here in 12 years, and we'll provide you with an external audit through NAEP every other year, but we want 100% of your students at proficiency or higher." In the meantime, states should take advantage of the specific ambiguity in the law and build the system that works best for them.

---

[5] *No Child Left Behind Act*, P.L. 107-110, Section 1111 (l), 107th Congress, 1st Session, 2001.

Building a System: Norm- vs. Criterion-Referencing

It is likely that the goals of AYP will be realized in ways that have not been pursued on a national basis, but which will be diligently pursued in individual states. Therefore, we would advise caution when overseeing developing systems, and not hasten to declare them insufficient in process so long as the outcome data they seek and produce match the goals and objectives of the law. Remember, this is about results, not process.

Accountability systems are still a new science. Few have been well researched. Many exist on paper, though few have been employed over any significant period of time. For this reason, educators, testing directors, and federal officials engaged in "approving" a given approach would be well advised to gather all of the pertinent data currently available. We may be in for a few surprises.

As an example, we hear a compelling and well-reasoned argument that the best method for testing students is to use a criterion-referenced test that has been tailor-made to directly correlate to a state's specific standards. If that argument is universalized as a compliance requirement of NCLB, every state that has not yet done so must commission the development of a specialized criterion-referenced test for use every year, rather than use any number of pre-existing commercial tests.

The argument for this approach says that only tests designed specifically around a state's standards can adequately reflect student progress toward those standards. Or so current accountability theory seems to suggest.

Theory is one thing, but we may miss potentially powerful state approaches if this theory dictates all future practice. In fact, requiring each state to develop an annual criterion-referenced test will immediately undermine extensive efforts already underway in states such as California, Arizona, and Tennessee, among others. These states currently use norm-referenced tests or test items to gauge academic progress down to the level of an individual student, and what they have found bears further study.

Some of their preliminary data suggest that this method of analyzing student achievement results in data comparable in quality and result to that derived from analysis of criterion-referenced tests. Until there is sufficient research in this area by those who know testing systems best, we should avoid dismissing the use of norm-referenced tests at the outset of this endeavor.

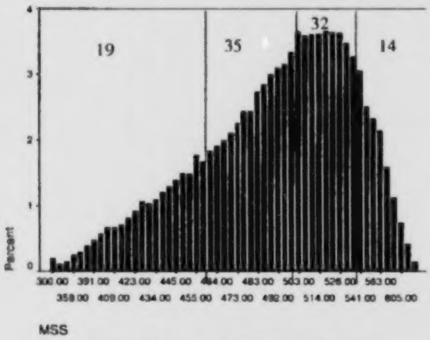
A quick look at Arizona's testing data should show why. Arizona administers both a criterion-referenced test (the AIMS test, shown in the left column on the next page) and a norm-referenced test (SAT-9, in the right column). If we lay the results of these two tests next to each other—understanding that there are technical differences in the administration of the tests that make a perfect correlation impossible—the results are still remarkably similar.[6]

---

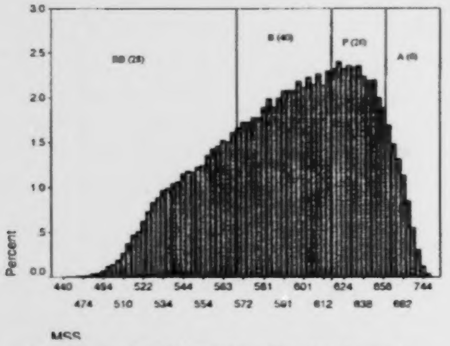
[6] In this particular case, percentile scores have been converted to normal curve equivalents for a more valid comparison of criterion- and norm-referenced test scores. (See above explanation in text.)

Figure 1. Results from Arizona's criterion-referenced test (on the left) and norm-referenced test (on the right) are remarkably similar.

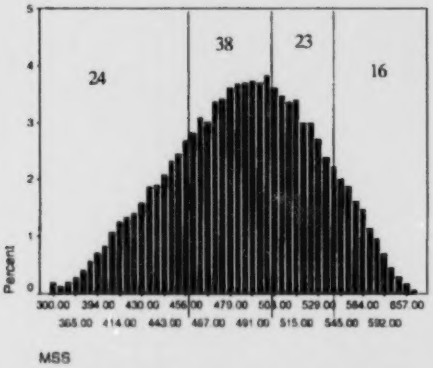
AIMS GRADE 3 MATHEMATICS



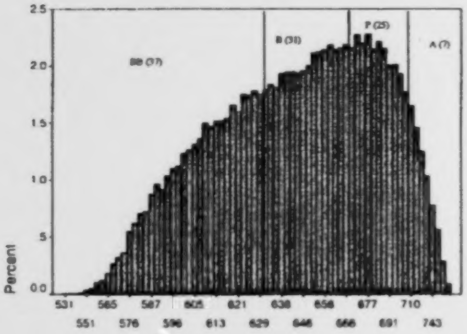
Grade 3 Mathematics Scale Scores Stanford 9



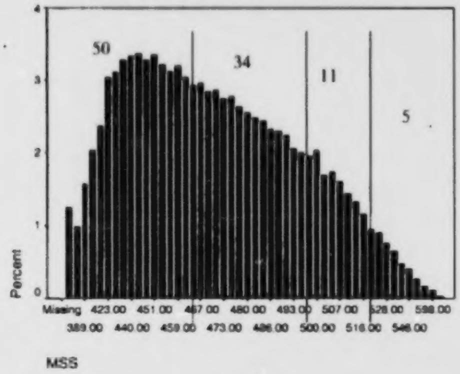
AIMS GRADE 5 MATHEMATICS



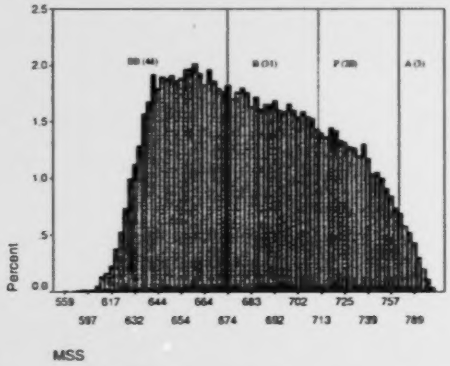
Grade 5 Mathematics Scale Scores Stanford 9



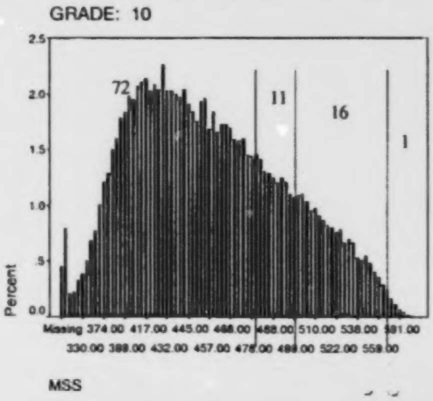
AIMS GRADE 8 MATHEMATICS



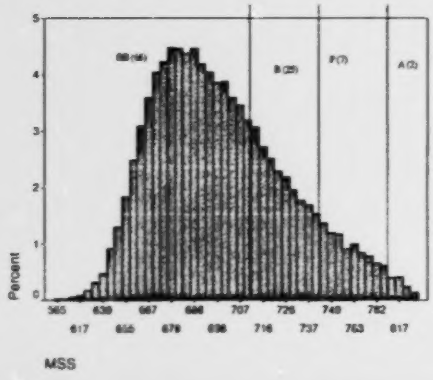
Grade 8 Mathematics Scale Scores STANFORD 9



AIMS GRADE 10 MATHEMATICS



Grade 10 Mathematics Scale Scores Stanford 9



It can, of course, be argued that a criterion-referenced test is more precisely matched to the state's specific standards. We don't disagree. Yet, norm-referenced tests are also based on a publicized set of standards, and these are generally consistent with those used for criterion-referenced tests. Bear in mind the goal of showing *progress*—a gain in knowledge of material deemed most essential for student success. Both a criterion-referenced and a norm-referenced test are made up of questions designed to make an effective judgment of student knowledge and skills in defined areas. Where they differ most significantly is presumably in their range of difficulty.

While a norm-referenced test seeks questions chosen to elicit a bell-shaped performance curve, the criterion-referenced test is made up of questions meant to match the standard. For norm-referenced tests, results are displayed primarily in a percentile ranking scale for comparison to other students, based on a nationwide "norming" population. However, most national norm-referenced tests also offer conversion of their percentile scores into a curve representing points given for every correct answer. As the Arizona data show, curves and performance levels for the converted norm-referenced tests nearly mirror criterion-referenced test results.

An additional point bears mentioning. Based on his work in Tennessee over the past 15 years, Dr. William Sanders offers the opinion that we do *not* need to have an excruciatingly tight match of state standards to specific test items. In fact, he places far more importance on "freshening" a test annually with new items than he does on specific linking to a particular standard.[7] It could well be that we have placed too much emphasis on states writing their own unique tests. This is yet another assertion that deserves additional study.

We are not arguing that criterion-referenced tests and norm-referenced tests are interchangeable. They are designed for different purposes and with distinct strengths and weaknesses, but the assumption that a state-developed criterion-referenced test better identifies student growth than a norm-referenced "test off the shelf" may not withstand in-depth analysis. The data produced by both norm- and criterion-referenced tests are so strikingly similar that an automatic preference for use of a criterion-referenced test to gauge student progress as part of NCLB seems unwarranted for the moment.

A final word in this regard: Those of us who support NCLB clearly believe that the core set of knowledge we seek for our students is sufficiently similar as to be assessable with a more generalized examination—otherwise, why the prominent role of the National Assessment of Educational Progress (NAEP) as an external audit for states in the new law? One cannot argue that gain can only be viewed within the confines of unique state assessments while simultaneously extolling the ability of NAEP to judge achievement across the board.

---

[7] Education Commission of the States. *A Closer Look: State Policy Trends in Three Key Areas of the Bush Education Plan—Testing, Accountability and School Choice*. (Denver: Education Commission of the States, 2001), 8.

The conclusion? We need more comparison and research regarding what these tests tell us. There are presently a number of states that not only use both norm- and criterion-referenced tests, but they also use them in different subjects, different grades, and, in some cases, in different locations around their state. Equating the results of this blend of norm- and criterion-referenced testing may be valid—and then again it may not. Until we have more data from the administration of these tests, and the opportunity to look at this data in a meaningful way, we ought not be in a hurry to junk the use of norm-referenced tests. Educators should currently worry less about whether a test is norm- or criterion-referenced, and concentrate instead on its relationship to state goals, and to collecting and analyzing the results of those tests in meaningful ways. We're looking at progress, not process.

High Stakes and Consequences

AYP requires states to disaggregate test results not only by communities and schools but also by specific sub-groups of students. Such disaggregation gives educators and parents a truer idea of what is really going on in their school—after all, a school that appears to be making progress when one looks at its average score may also show, upon closer examination, that certain groups of students have made little or no gains. Disaggregation of results is a necessary tool of accountability to ensure that schools do not hide failing groups of students behind the law of averages.

So, what happens if students in a school or in a particular subgroup do not meet or exceed the state's defined standard for AYP? The answer is simple: that school would not make adequate yearly progress. The NCLB is very clear about the consequences that such schools will face, and the stakes are high.

If schools and districts do not show gain over a defined period of time, action will be taken on behalf of the students in those schools, including mandatory public school choice and the provision of individual supplemental services purchased with Title I funds. In addition, chronically failing schools face the very real possibility of having their schools completely restructured, while states that fail to meet their obligations under their state plan risk the loss of federal administrative dollars.

These potential penalties resonate loudly with schools, districts and states, and they send a clear message to parents that the law is serious about providing them opportunities to remove their children from consistently-failing schools. In a welcome break with past policy, school failure will result in meaningful consequences, and will empower parents to immediately remove their children from failing schools, instead of consigning them to continued failure. Further, in a contrast to the overall mood of NCLB, the timelines and sanctions imposed for school failure are specific and non-negotiable, as they should be. There is simply no more room for flexibility when it comes to consequences for failing schools.

If a school fails to make adequate yearly progress for two consecutive years, it will be identified by the district and state as *needing improvement*. This identification will mean that federal funds will be available to states and districts to provide schools with technical assistance to improve academic achievement—but financial assistance alone is no longer seen as a sufficient tonic for the ailment. The school is also subject to stricter and more rigorous sanctions to ensure that change occurs as quickly as possible. After two years of failure, the district is required to create a plan to turn the school around and to offer public school choice to all students in the failing school by the beginning of the next school year. Further, the district must pay the costs of transporting any students who opt to attend a different public school, including public charter schools.

If a school fails to make adequate yearly progress for three consecutive years, it must not only continue to offer public school choice for all students, but must also allow disadvantaged students in the failing school to use Title I funds to pay for supplemental services from a provider of choice. Schools will be required to set aside 20 percent of their total Title I allocation to pay for both the supplemental services and transportation to these services. Not less than 5 percent must be used for each.

After four years of failure to make adequate yearly progress, districts are required by law to implement *corrective action* in their school. This means that, in addition to continuing the provision of public school choice and supplemental services, districts must intervene more forcefully. This could mean removing school staff, changing school leadership, or altering curriculum and programs. Finally, to stem the tide of continuous failure, any schools that fail to make adequate progress for five consecutive years would be completely restructured. This might mean a state takeover, alternative governance, private management, new staff, or becoming a charter school. In essence, they will begin anew.

Schools will be released from the "corrective action" category only after making adequate yearly progress for two consecutive years.

With the enactment of NCLB, these consequences go into immediate effect for schools that have already been identified as in need of improvement under the IASA. These schools—some 6,700 of them[8] —are considered to be in their first year of school improvement (in 2001-2002) and must offer public school choice in the coming school year (2002-2003). Likewise, the 3,000 schools that are already in their second year of school improvement under the previous law must provide individual student services to supplement the regular school day in addition to public school choice for all low-income students in the coming year. This means students who have been in schools identified as failing for two or three years will receive immediate help through NCLB. The clock does not start over for these students, and failing schools do not receive an amnesty period simply because the law changed.

Just as schools are held to showing results under the AYP process, so too are school districts and, ultimately, the state. The state, usually through its state department of education, is responsible for determining whether an LEA has made progress, and identifying whether it needs improvement or requires corrective action. Likewise, progress by the state toward meeting its AYP objectives is reviewed by the U.S. Department of Education, using a peer review process. States that do not have in place standards and assessments, a system for measuring and monitoring AYP, or a mechanism for publicly reporting results risk having their funding for state administration withheld.[9]

---

[8] House Committee on Education and the Workforce, *Press Release: H.R. 1 Education Reforms Would Mean Immediate New Options for Students In Thousands of Failing Schools—Beginning in 2002*, December 13, 2001.

Additionally, any State education departments that have been granted "flexibility authority" will lose that authority if the state fails to make adequate yearly progress for two consecutive years. Similarly, local education agencies that are participating in local flexibility demonstration projects would also lose that opportunity if their schools fail to make adequate yearly progress for two consecutive years.

While there are consequences for schools not meeting or exceeding the goals of adequate yearly progress, there are also rewards and recognition for schools that do make expected progress. Schools that significantly close the achievement gap or that exceed the AYP requirements can receive the State Academic Achievement Awards, and schools that make the greatest gains will be eligible for the Distinguished School Award. Along with the schoolwide recognition, teachers could receive financial awards in schools that receive the Academic Achievement Awards.

The Importance of Rolling Averages

In defining what is meant by AYP, we mentioned that states may use a three-year rolling average of their assessments. This is relevant because there has been some concern expressed about states placing too much emphasis on the most recent test scores and about how single-year scores exaggerate sometimes-random fluctuations that occur from one year to the next.[10] Therefore, the process outlined in NCLB allows states some flexibility regarding the establishment of a uniform averaging procedure by using data from one or two school years immediately preceding the current year, instead of just the scores from a single year.

For example, states beginning to define their AYP expectations will use 2001-2002 school year test scores. However, NCLB allows the states to average in scores from 2000-2001, as well as data from 1999-2000—the two preceding years. During 2002-2003 school year, the data from 2000-2001 and 2001-2002 would be used in computing for the school's average, while the 1999-2000 data would be dropped, thus establishing a three-year rolling average. Each year, then, the rolling average will incorporate the current year and the two previous years.[11]

Why is this important? As the system moves forward and multiple years of data become available, the reliability will be increased. Certainly, schools that do not have scores from previous years will be at a disadvantage, and results from new schools will be more volatile and less reliable until they can establish at least three years of data and begin the rolling average.

---

[9] *No Child Left Behind Act*, P.L. 107-110, Section 1111 (g)(2), 107th Congress, 1st Session, 2001.
[10] Kane, 10.
[11] *No Child Left Behind Act*, P.L. 107-110, Section 1111 (J), 107th Congress, 1st Session, 2001.

It is also important to note that, after establishing a baseline of student achievement using the 2001-2002 data, states are given the opportunity to confirm the results during the following year. The confirmation of this year of data means that schools, districts, or states that have not been currently identified for school improvement would not automatically be considered as in need of improvement based on a single year's worth of data.

An Exercise in Humility

Today, there is no obvious template or ideal model that states can turn to in the development of their AYP process. Experience is too brief, research too new, and approaches too varied to yet have yielded a definitive prototype—but the experimental nature of the process is part of what makes it both intriguing and worthwhile. We need education leaders who are not afraid to experiment, who are open minded about varying approaches to assessment, who are research oriented, and who have a sincere desire to learn what really works before rushing to declare that an ideal model has been found. What is really called for is humility.

This will be an exercise in humility for all parties involved in the process. Education leaders in the nation who have created, enacted, or lived with a particular approach to assessing student gain over time must share their own experience and be willing to accept approaches they may not have considered or even discarded.

There remains at the core of NCLB, however, a set of non-negotiable principles and requirements based on the experience and wisdom of these same leaders. The law outlines for states a highly desirable accountability infrastructure that is stringent in and of itself—and presumably sufficient to produce desired results, when applied in tandem with improvements in instruction, curriculum, and high expectations.

We will do well to recall the work of many states and leaders in the preceding decade that has brought us what knowledge we currently claim in this arena. That knowledge is yet young and still evolving. We should focus on meeting the major goals and let the science of accountability evolve.

This bill enacts a new vision of American education. Its goals are idealistic, and they are achievable if we are to believe the work going on in hundreds of school across the nation today. "No Child Left Behind" now means just that. Whether states can attain that goal is yet to be seen—but the gauntlet has been thrown down, and we should pick it up. America's children are waiting for us to meet the challenge.

References

Bush, George W. *No Child Left Behind*

Education Commission of the States. *A Closer Look: State Policy Trends in Three Key Areas of the Bush Education Plan – Testing, Accountability and School Choice.* Denver: Education Commission of the States, 2001.

Education Commission of the States. *Building on Progress: How Ready Are States to Implement President bush's education Plan?* Denver: Education Commission of the States, 2001.

House Education & the Workforce Committee. *Press Release: H.R. 1 Education Reforms Would Mean Immediate New Options for Students in Thousands of Failing Public Schools – Beginning 2002*, December 13, 2001.

The No Child Left Behind Act of 2002, Public Law107-10, 107th Congress, 1st Session 2002.

House Education & the Workforce Committee. *Fact Sheet: Bush Testing Plan Measures Results, Empowers Parents.* <http://edworkforce.house.gov/issues/107th/education/nclb/factbushtest.pdf>.

House Education & the Workforce Committee. *Fact Sheet: H.R.1 Conference Report Highlights: Accountability for Student Achievement.* <http://edworkforce.house.gov/issues/107th/education/nclb/accountfact.htm> December 10, 2001.

House Education & the Workforce Committee. *Fact Sheet: H.R.1 Conference Report Highlights: State and Local Flexibility.* <http://edworkforce.house.gov/issues/107th/education/nclb/statelocalflex.htm> December 10, 2001.

House Education & the Workforce Committee. *Talking Points: What the H.R.1 Education Reforms Mean for States.* <http://edworkforce.house.gov/issues/107th/education/nclb/tpsstates.htm>, December 10, 2001.

Kane, Thomas J., et al. "Assessing the Definition of 'Adequate yearly Progress' in the House and Senate Education Bills." July 15, 2001.

LeTendre, Mary Jean. " Defining Adequate Yearly Progress: Strengthening Responsibility for Results Without Toppling State Accountability Systems," <http://www.ctredpol.org/pubs/LeTendreFinalPaperAYP.pdf>.

U.S. Department of Education. "No Child Left Behind: Achieving Equality through High Standards and Accountability." <http://www.ed.gov/inits/nclb/part3.html>, August 21. 2001.

No Child Left Behind: Who Is Included In New Federal Accountability Requirements?

Richard J. Wenning, Paul A. Herdman, Nelson Smith

INTRODUCTION

"Leave no child behind." Powerful in its simplicity, daunting in its complexity, this is the challenge posed by the President and Congress in reauthorizing the Elementary and Secondary Education Act (ESEA). The legislation seeks to make good on its promise through a substantial expansion of the federal role in education, particularly in the area of accountability. This paper reviews how the legislation will operate with respect to different groups of students and schools, and examines factors that could delay or dilute its guarantee of educational accountability for the academic achievement of all children.

As standardized testing has expanded, so has the list of well-intentioned arguments for excusing low achievement by whole categories of students. While special education law provides for testing with "accommodations," in practice it has pushed educators to focus more on procedural compliance than student outcomes. The achievement of language-minority students has often been overlooked or mismeasured as school districts lacked the skill or will to administer appropriate assessments. State laws have required charter schools to participate in statewide testing, but have largely treated accountability reporting as an afterthought.

The new law – the No Child Left Behind Act of 2001 (NCLB) – appears to mean business in all these cases: Its title leaves no room for ambiguity and, in a major expansion of the federal role, the Act requires annual testing; specifies a method for judging school effectiveness; sets a timeline for progress; and establishes a sequence of specific consequences in the case of failure. This paper examines four questions that will help determine whether the new law's ambitions will be achieved:

- What kinds of tests must be used and when?
- What students must take the tests and who is exempted?
- Whose scores count and how must they be reported?
- How do the Act's testing and Adequate Yearly Progress (AYP) requirements apply to different kinds of schools, including private schools, home schools, and charter schools?

The paper is divided into four sections. The first provides context on the law, its intent, and its implementation to date. The second section focuses on students, examining who gets tested and when and whose scores "count" for accountability purposes, with a particular focus on students with special needs. The third section focuses on the measurement of school performance and the applicability of accountability provisions to private, charter, and home schools. The final section offers conclusions and recommendations for policymakers.

THE EBB AND FLOW OF ACCOUNTABILITY REQUIREMENTS: NCLB IN THE CONTEXT OF THE LAST TWO ESEA REAUTHORIZATIONS

In order to understand how the law will affect students, it is important to understand its purpose and how it has evolved. The general intent of the ESEA has remained relatively unchanged since its enactment in 1965:

> *To ensure equal educational opportunity for all children regardless of socioeconomic background and to close the achievement gap between poor and affluent children by providing additional resources for schools serving disadvantaged students.*[1]

While the ends have remained constant, the means for measuring progress have changed over time. The 1988 reauthorization of ESEA established a new accountability system for Title I (then Chapter 1). Its Program Improvement provisions required local education agencies (LEAs) to identify schools with ineffective Chapter 1 programs on the basis of average individual student gains on annual standardized, norm-referenced tests, and to provide capacity-building support. While the Department of Education encouraged districts to establish additional desired outcomes, to be measured by criterion-referenced tests or other indicators, most stayed with the default option: average annual gains on norm-referenced tests.

The 1994 reauthorization of ESEA, the Improving America's Schools Act (IASA), reflected the national momentum toward standards-based reform. It also dealt with increasing concerns about reliance on a single test, including the likelihood that many schools were judged effective or ineffective on the basis of changes in test scores that were due to random fluctuations.[2] The IASA accountability provisions:

- Eliminated the annual testing requirement and replaced it with a requirement for testing in three grades (at least once within each of the following grade spans: 3-5, 6-9, and 10-12).
- Included a requirement that test scores be disaggregated by multiple categories, e.g., race, language proficiency and disability.
- Removed federal guidelines for measuring annual school performance and minimum progress, instead requiring each state to define "how good is good

---

[1] Generally, Public Law 107-110, section 1001.

[2] General Accounting Office, "Chapter 1 Accountability: Greater Emphasis on Program Goals Needed" (GAO/HRD-93-69, 1993).

enough" in terms of a school's Adequate Yearly Progress (AYP), resulting in many different approaches among the several states

In sum, IASA encouraged each state to create a coherent system of standards and accountability rather than a separate system for Title I students, while at the same time, allowing substantial variation from state to state. It also marked a departure from annual testing, thereby removing the federal incentive to track student progress over consecutive years. NCLB merges elements of the two prior reauthorizations: restoring the annual testing obligation of 1988 and retaining the standards-based emphasis of 1994.

NO CHILD UNTESTED? WHAT TESTS WHEN; WHO GETS TESTED; AND WHOSE SCORES COUNT?

NCLB expands federal testing requirements, eventually mandating annual testing for all public school students, but does not necessarily make all students' performance "count" for school accountability purposes.

What Tests When?

Left to their own discretion, states have created a broad array of approaches to measuring student performance. Some states test reading and math every year; others test only those subjects at three or four-year intervals, and others test a variety of subjects in a variety of grades.

One critical difference is whether states use *norm-referenced* or *criterion-referenced* tests (some favor one or the other, and some use both). Advocates of standards-based reform prefer the criterion-referenced variety because they can be directly aligned to a given state's standards. However, precisely because they are generally custom-fit for each state, they are far more expensive to create and produce results that are more difficult to compare from state to state.

An Emphasis on Criterion-Referenced State Tests. Like the 1994 law, NCLB encourages states to develop criterion- versus norm-referenced tests. The legislation requires that assessments be aligned to states' content and student academic achievement standards, and that states define benchmarks of proficiency. However, while the Act mandates annual testing by 2005-2006, it does not explicitly require states to administer the *same* test from year to year. Thus, states like Louisiana and Maryland that test students in grades three through eight with a mix of norm- and criterion-referenced tests may technically be in compliance, yet produce results that lack consistency over time. This arrangement may not pass federal muster, however, when states are asked to demonstrate AYP.

As to what subjects are tested, and when, states have some flexibility, particularly early on. Prior to 2005-2006, they must measure proficiency of mathematics and reading or language arts, and, as under the 1994 requirements, do this at least once during grades three through five, six through nine, and 10 through 12. Beginning in 2007-2008, states must also include science assessments at least once during each of these three grade spans. By 2005-2006, states must measure student achievement annually against state academic content and achievement standards in grades three through eight in mathematics and reading or language arts. So, by 2007, students will be tested annually from grades 3 to 8 in reading and math, tested twice in the elementary grades in science, and then in reading, math, and science at least once in grades 10-12. (States may also choose to add other subjects into the testing mix.)

Since definitions of "proficiency" can vary dramatically from state to state, beginning in the 2002-2003 school year, every state must also participate in biennial assessments of fourth- and eighth-grade reading and mathematics under the National Assessment of Education Progress (NAEP) – at least so long as Congress appropriates funds to underwrite such assessments.

Who Gets Tested?

NCLB extends federally mandated testing to a greater proportion of students than ever before by reaching all student groups, not just those served by Title I. Its testing requirements cover all public elementary and secondary education students, including students attending charter schools. As provided for under Section 1111(b) (3) (C) (i) of Title I, these assessments must "be the same academic assessments used to measure the achievement of all children." Further, state assessments must be disaggregated within each state, LEA, and school by student demographic subgroups, including:

- economically disadvantaged students;
- students with disabilities;
- students with limited English proficiency;
- major racial and ethnic groups; and
- gender

This provision attempts to rectify distortions and variations masked by the widespread reliance on schoolwide averages. For example, schools discovered that they could run up average test scores by allowing a liberal-leave policy for low achievers on test day. And districts found that they could garner good press by steering resources to high-achievers who could boost average test scores. NCLB addresses both problems by insisting that fully 95 percent of students be tested and tying incentives to the performance of disaggregated student groups.

This is cause for real celebration in the case of students with disabilities and those with limited English proficiency (LEP), segments of the national student population too often subject to what President Bush has called "the soft bigotry of low expectations." In the past, when states were given the discretion to make their own exemption decisions, the result was widespread exclusion of students with disabilities from large-scale state and national assessments. Indeed, as recently as 1995, a review of state and national data collection programs found that, at the national level, 40 to 50 percent of school-age students with disabilities were estimated to be excluded from the most prominent national education data collection programs (e.g., National Assessment of Educational Progress).[3]

Reasons for such exemptions ranged from a desire to protect students with disabilities from the stresses of testing, to a lack of awareness of the availability of test modifications or accommodations, to an aversion to the difficulties of specialized test administration, to the desire to raise a school's average scores.[4] Whatever the impetus, the results were personally damaging not only to the many students improperly impeded from achieving and stigmatized by exclusion, but also to reform efforts in general. If students with disabilities do not participate in testing, there is no performance data to assess and therefore they cannot be meaningfully included in any resulting systemic reform. They get left behind.

Limited English proficient students with disabilities present a particularly complex set of problems, because language complicates the process of identifying their disability. Districts fearing misdiagnoses because of a language barrier may allow such students to remain in English as a Second Language (ESL) or other transitional classes for the maximum three years allowed under most state laws before they are assessed. Of the nation's 2.9 million students enrolled in programs for English Language learners, an estimated 184,000 have disabilities, according to the U.S. Department of Education.[5] NCLB's provisions to clarify the time frame for participation in ESL tracks, coupled with the expectation for 95 percent participation within student subgroups, should serve to mitigate this problem.

In any case, the good news is that NCLB unmistakably includes both students with disabilities and LEP students under its testing and accountability provisions, and reinforces prior federal requirements for reasonable accommodations needed to achieve that end. (Of course, the interpretation of "reasonable" remains subject to wide discretion and no one should expect rancorous disputes and lawsuits on this point to taper off.)

In the case of LEP students, the legislation goes so far as to require testing in English proficiency beginning in the 2002-2003 school year. This is a major departure from the 1994 law, and a clear signal of federal intent that achievement standards should apply to all students—and that everyone should become proficient in English.

Wisely, the bill's framers included a safety catch to ensure statistical significance and protect the identities of individual students when disaggregation creates very small student groups. For the purposes of determining Adequate Yearly Progress, or "AYP," such disaggregation "shall not be required in a case in which the number of students in a category is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student." This language is also used under Sec. 111(b) (3), which sets forth the requirements of state assessments.

---

[3] See McGrew, Kevin, et al., "Why We Can't Say Much About the Status of Students With Disabilities During Educational Reform," NCEO Synthesis Report No. 21, National Center On Educational Outcomes, August 1995. Available at coled.umn.edu/NCEO/OnlinePubs/SynthesisReport21.htm. Inclusion rates varied significantly by state. Ibid.

[4] See Heubert, J.P. and Hauser, R.M., (Editors). (1998). "High Stakes: Testing for Tracking, Promotion and Graduation, Washington D.C.:" National Research Council, p. 193.

[5] Mary Ann Zehr. "Bilingual Students with Disabilities Get Special Help." Education Week: 7 November 2001.

It is unclear whether states, districts, or individual schools will have the final decision about whom to test (or not to test). The likely scenario will be that states will define the requirements and accommodations for state testing and districts and schools will be charged with implementing those guidelines faithfully. As this is addressed as part of the U.S. Department of Education's regulatory process, it is likely that the pre-existing civil rights laws governing special populations of students will drive the debate.

Whose Scores Count and How Must They be Reported?

Adequate Yearly Progress. While substantially all students must participate in state testing programs, not all students' scores will necessarily count equally in the alignment of incentives for improving school performance. The key question is whether scores are included in measuring "Adequate Yearly Progress," or AYP. NCLB provides a new federal definition of AYP that is more specific than the 1994 reauthorization while still preserving some state latitude:

- Each state, using data from the 2001-2002 school year, must establish a starting point for measuring the percentage of students meeting or exceeding the state's proficient level of academic achievement on the state assessments.[6]
- States must develop a 12-year timeline in which all students, within each of the "disaggregated" subgroups, will attain proficiency on the state assessments.
- States must develop annual measurable objectives that are consistent across schools and student subgroups and increase in equal increments over 12 years, with the first increase required to occur in not more than two years, and the remaining increases to occur in not more than every three years.
- States may establish a uniform procedure for averaging data over multiple years and across grades in a school.

The Act prescribes far more extensive consequences for failure to achieve AYP than in previous reauthorizations. However, unlike the universal testing requirement, which applies to all schools, those sanctions apply only to schools that receive funds under Title I.

Reporting results. The legislation's public-accountability provisions are impressive. Beginning in the 2002-2003 school year, states must provide parents and the public with annual report cards, which include information on student achievement disaggregated by race, ethnicity, gender, disability status, English proficiency, socioeconomic status, and migrant status.

---

[6] In establishing this starting point, the state must use the higher of either the proficiency level of the state's lowest-achieving group or the proficiency level of the students in the school at the 20th percentile in the state, among all schools ranked by the percentage of students at the proficient level.

Taken together, the AYP and reporting provisions provide a new level of transparency about school performance[7], enabling parents, administrators, and public officials to make accountability more than a slogan. Yet a closer look reveals two potentially significant concerns:

First, since grade-level performance does not need to be monitored, schools can provide school-wide averages across grades rather than reports for all student subgroups in each grade. This makes sense; the matrix required to present every subgroup in every grade would be unwieldy. Yet without such reporting, schools can focus their energies on grades with higher achieving students -while ignoring grades with lower achieving students – and still increase their school average.

Second, and perhaps more serious is NCLB's perpetuation of the Law of Averages: making the schoolwide average of student proficiency the basic yardstick of progress. Although results will be disaggregated by student subgroups, reliance on this measure may discourage use of "value-added" analytical methods, which measure the impact of a school on the progress of individual students over time. States, however, have latitude in this area and there is reason for hope that such analytical methods will be used given that the NCLB provides permission and financial incentives for states to use such methods. The Act (in Title I, Part A, Section 1111, subsection 3B) states that: "*Each State educational agency may incorporate the data from the assessments under this paragraph into a State-developed longitudinal data system that links student test scores, length of enrollment, and graduation records over time.*" The Act also authorizes federal funding for states interested in developing longitudinally linked student databases (TitleVI, Part A, Section 6111).

Nevertheless, because the new federal definition of AYP encourages the analysis of *average* proficiency levels across student groups, the progress of *individual students* could be lost. While a problem for state and national policymakers, this weakness in the Act may undermine its utility most seriously at the school and district level. When there is no annual measurement of individual student performance over time, educators lack important data needed to evaluate their own work – to understand the "value added" by their efforts. Without student-level results, administrators can face chaos in evaluating the impact of teachers and schools. This is especially true when there is high student mobility (as in many urban systems), or in the case of newer charter schools, when entire grades of students are added from year to year. Comparisons of schoolwide averages can be misleading and uninformative when the composition of classes changes so dramatically from one year to the next.

Arguably, the measurement of progress required by NCLB confuses the building for the kids. Without a focus on student progress over time, superintendents and state boards of education will be measuring the percentage of students at the proficient level and calculating the change from year to year – but the numbers will refer to the apples who were in the building last year versus the oranges there now. Judgments about school performance may have little to do with how a given cohort of students is actually affected by their schooling over time.

---

[7] It should be noted that Section 1116© also provides for LEAs to be identified as in need of improvement.

Implementation and Enforcement Matter. While the rhetoric of inclusion is promising, it will ring hollow if the bill is implemented poorly. The state and federal record on this issue is not encouraging. A Department of Education study of Title I, released seven years after the passage of IASA, found that, of the 34 states reviewed, 13 did not have adequate testing and accountability provisions for limited English proficient students; 10 had similar difficulties with disabled students; and 16 had difficulty in disaggregating the data as required.[8] Moreover, while few states have met the requirements of IASA even now, no state education agencies have been financially penalized for not complying with ESEA.[9]

If no child is to be left behind, states will have to meet a significant implementation challenge and the federal government will have to think anew about its own enforcement role. Traditionally, the federal role has been top-down and compliance-driven, a combination of Bad Cop and Federal Nanny. For example, the 1997 amendments of the Individuals with Disabilities Education Act (IDEA) paid lip service to outcomes-oriented accountability, but the Department of Education's regulations reverted to form. Commenting on the Department's enforcement system, analysts Patrick J. Wolf and Bryan C. Hassel said it is "flawed in design because, instead of replacing a rules-driven oversight process with a results-driven oversight system, it instead merely piles more rules regarding performance assessment into the previous process-based compliance system which remains largely intact but overwhelmed with paperwork."[10]

Among the mechanisms that might be explored to reach NCLB's inclusion goals are highly publicized annual rankings of how well states do in testing all subgroups; setting timelines with goals for improvement rather than the existing (rather mild) sanctions for failure; withholding only administrative funds rather than those that go to schools; and convening multi-state panels to help struggling states address technical problems.

---

[8] U.S. Department of Education, "High Standards for All Students: A Report from the National Assessment of Title I on Progress and Challenges Since the 1994 Reauthorization" (January 2001).

[9] Robelen, Erik W., "States Sluggish on Execution of 1994 ESEA." Education Week 28 November 2001. <www.edweek.com/ew/newstory.cfm?slug=13comply.h21>.

[10] Bryan C. Hassel and Patrick J. Wolf, "Effectiveness and Accountability in Special Education (Part 2): Alternatives to the Compliance Model." In Chester E. Finn, Jr., Andrew J. Rotherham, and Charles R. Hokanson, Jr.,Eds. Rethinking Special Education for a New Century. Washington, DC: Thomas B. Fordham Foundation and Progressive Policy Institute, 2001: 309-334. Available: http://www.edexcellence.net/library/special_ed/special_ed_ch14.pdf.

APPLICABILITY OF NCLB ACCOUNTABILITY REQUIREMENTS TO DIFFERENT KINDS OF SCHOOLS

NCLB gives special consideration to private schools, home schools, and charter schools. In the case of charter schools, the Act presents some real challenges, as well as some latitude, for their accountability relationships with their sponsoring agencies.

Applicability to Private Schools and Home Schools

The testing and AYP requirements of the NCLB apply only to private schools (and then only to specific students) that receive funds or services under the Act. In contrast, home schools are totally exempted from the Act's provisions Section 9506 of the Act, pertaining to private, religious, and home schools, provides the following:

> " *(a) Applicability to Nonrecipient Private Schools.--Nothing in this Act shall be construed to affect any private school that does not receive funds or services under this Act, nor shall any student who attends a private school that does not receive funds or services under this Act be required to participate in any assessment referenced in this Act.*
>
> " *(b) Applicability to Home Schools.--Nothing in this Act shall be construed to affect a home school, whether or not a home school is treated as a home school or a private school under State law, nor shall any student schooled at home be required to participate in any assessment referenced in this Act.*
>
> " *(c) Rule of Construction on Prohibition of Federal Control Over Nonpublic Schools.--Nothing in this Act shall be construed to permit, allow, encourage, or authorize any Federal control over any aspect of any private, religious, or home school, whether or not a home school is treated as a private school or home school under State law. This section shall not be construed to bar private, religious, or home schools from participation in programs or services under this Act.*
>
> " *(d) Rule of Construction on State and Local Educational Agency Mandates.--Nothing in this Act shall be construed to require any State educational agency or local educational agency that receives funds under this Act to mandate, direct, or control the curriculum of a private or home school, regardless or whether or not a home school is treated as a private school under State law, nor shall any funds under this Act be used for this purpose.*

Funding of private-school programs must be on an equitable basis with all other children receiving Title I assistance. The LEA is required to consult with private school officials to determine how children's needs will be identified and what services will be offered; these can be provided either directly by the LEA, or through contracts with "public and private agencies, organizations and institutions." With respect to testing, the consultation must cover "how the services will be academically assessed and how the results of that assessment will be used to improve those services." Private schools are given an explicit process of complaint to the state education agency if they believe the consultative process was not "meaningful and timely," but the state agency is provided no guidance on what sort of remedy to concoct.

The private school provisions seek to create the same incentives for testing and improvement as will exist for public schools, but stop well short of spelling out clear consequences in deference to the established tradition of federal noninterference in the curricula of private schools.

Applicability to Charter Schools

As public schools, charter schools are subject to the Act's testing and AYP requirements; however, specific language acknowledges their status as autonomous public schools operating under performance agreements with the agencies that authorize their charters, often referred to as *authorizers*. Depending on state laws, charter school authorizers may include state boards of education, colleges and universities, municipal bodies, special-purpose agencies, and most commonly, local school districts.

The legislation raises important questions about how charter schools should fit into the larger scheme of federal accountability requirements, because state laws grant authorizers the authority and responsibility to oversee and evaluate charter school performance and accountability according to measures set forth in their charter agreements. Because some authorizers are not local or state education agencies – those being the agencies forming the regulatory structure of NCLB – the legislation could potentially create confusion and redundancy in oversight roles or accountability requirements for charter schools.

To avoid such confusion, the NCLB maintains traditional federal deference to state law, stating simply that, "The accountability provisions under this Act shall be overseen for charter schools in accordance with State charter school law." The following report language amplifies the legislative intent:

> *"Charter schools are public schools and therefore subject to the same accountability requirements of this Act as they apply to other public schools, including Sections 1111 and 1116, as developed in each state. However, there is no intent to replace or duplicate the role of authorized chartering agencies, as established under each state's charter school law, in overseeing the Act's accountability requirements for the charter schools that they authorize. Authorized chartering agencies should be held accountable for carrying out their oversight responsibilities as determined by each state through its charter school law and other applicable state laws. This should be done in ways that do not inhibit or discourage the approval or oversight of innovative, high quality charter schools."*

Implementing this approach will take some doing. Given the wide variety of charter founding groups and school missions, authorizers and state officials face complex judgments in weighing these new federal provisions against existing federal and state laws, and against the charter contracts already executed. Areas of potential conflict include:

Aligning Timelines for Corrective Action. Authorizers will need to examine how charter school renewal decisions, which occur every three to five years in most states, will align – or perhaps clash -- with the federal timelines for improvement, which require states to denote equal annualized improvements over a 12-year period. For example, if a pre-existing charter school has a five-year charter and its state test scores warrant corrective action in year two under the Act's accountability provisions, what takes precedence?

Charter-Specific vs. State-Mandated Objectives. Authorizers will need to decide how to weigh a school's performance on charter-specific goals against its performance on a given state's test. If Public Service Charter School is meeting its objective of teaching life skills through service learning, but students are behind the state benchmark in mathematics how should the charter specific goals figure into accountability decisions?

Special Populations. Many charter schools go into business expressly to serve "at-risk" populations. Is it fair to apply AYP to a school serving a population of recent immigrants speaking Creole or Farsi? That school's charter may set forth a pedagogically sound timeline for student attainment of English proficiency, but it may not match the AYP framework. (In fact, the same point could be made about many traditional public schools as well.) And what about unusual learning settings such as "virtual" or "independent-study" charter schools?

Starting the Clock on Charter Schools and Applying Corrective Actions. Most charter schools are still in their startup-stage, with roughly half of all charter schools less than four years old. Since all new schools need time to get established, it may make sense to assess baseline performance levels after a school's first or second year. Authorizers will need to decide how much of a grace period is permissible, and when the "clock" for corrective action should start.

The Act also provides for a host of corrective actions that pose challenges for charter schools and their authorizers. In general, these corrective actions were designed with traditional schools and their districts in mind, not charter schools that may be treated as LEAs (the traditional designation of a school district) by states for grant purposes under some charter school laws. For example, in the 2002-03 school year, the Act provides for corrective actions for schools not meeting AYP that include, among other things, requiring the LEA to allow students attending such a school to choose another public school and for the school to develop an improvement plan to address AYP problems, as well as provide assurance that 10 percent of Title I funds will go toward professional development. The Act also requires LEAs to provide (or pay for) transportation and to use up to 5 percent of their Title I-A funds for such purpose. If a school again fails to make AYP, the LEA must, among other things, continue to provide public school choice and use a prescribed portion of Title I funds to pay for supplemental services or transportation.

Each year a school fails to achieve AYP, corrective actions escalate, culminating in reconstitution or outsourcing the school's management. Not only may these corrective actions may be at odds with existing accountability agreements between charter schools and their authorizers, but the requirements for LEAs to fund specific remedies may fall on charter schools with LEA status or their authorizers (if the authorizer is itself an LEA).

It appears that states and charter school authorizers have considerable discretion in answering the questions and concerns raised above. Nevertheless, some of these issues will no doubt receive attention during the Education Department's regulatory process. As the implementation and regulatory processes unfold, it will be important to allow the accountability relationships between charter schools and their authorizers to develop without undue encumbrance. The quest for flexibility in these arrangements should not be viewed as an effort to evade accountability.

Rather, charter schools seek to find many paths to the same high standards sought for all other public schools. Under the new legislation, they may powerfully demonstrate the idea advanced by Paul Hill that setting fixed outcomes might serve to free schools to explore unique approaches to meeting those goals. Hill argues that, when we define public education as "a commitment to a goal of universal competency rather than a fixed set of institutions," we are required to continually search for the best way to educate children and open ourselves to the possibility that "any locality might pursue many different approaches."[11] It is possible that traditional school districts may learn a great deal from watching how charter schools use their freedom to pursue the new accountability goals. At the same time, we hope the law will not stifle charter schools' pursuit of success according to school-tailored measures beyond state-mandated AYP, as such other measures can also be greatly instructive for conventional school systems.

CONCLUSIONS

The No Child Left Behind Act of 2001 is a solid step in the direction of establishing a new nationwide commitment to the high academic achievement of all children. It is also underwritten by a bold expansion of the federal role in education.

The most obvious conclusion is that the law must be implemented well. The Department of Education should study and report in plain language on how well states and school districts fulfill these responsibilities. Special notice should be given to the provisions setting new test-taking targets, to ensure that the commendable goal of testing at least 95 percent of students is met and does not result in leaving behind the five percent most in need. In short, the key challenge for policymakers (at all levels of government) in refining the NCLB will be in developing enforceable incentives without overburdening school leaders, while simultaneously ensuring that schools have the resources they will need to succeed.

---

[11] Hill, Paul T., "What is Public About Public Schooling?." in Terry M. Moe, ed., A Primer on America's Schools, Palo Alto, CA: Hoover Institution, 2001.

A second imperative is to include a variety of stakeholders in the regulatory and enforcement processes, reflecting the myriad ways we now educate students. It is especially important that charter schools and authorizers be given the opportunity to create and demonstrate sound oversight regimes that follow federal policy while respecting state laws. One of the most promising educational reforms in decades should not be stifled by a bureaucratic, one-size-fits all approach to federal regulation.

Finally, the Department of Education should move to expand and strengthen the quality of data collected for accountability purposes. The measures contained in NCLB are not bad ones; indeed, they are an improvement over prior accountability schemes. By mandating annual testing of entire school populations, they create an opportunity, but not an obligation, to measure of the progress made by cohorts of students over time. The Department of Education will have ample opportunity through the regulatory process to signal its support for states' and districts' use of such "value-added" measures of school performance. Congress should back this up with enough funds so the Secretary can make grants to states that wish to develop longitudinal data systems.

After years of worry over what might happen in this round of ESEA reauthorization, and after months of horse-trading in which no side got all it wanted, Congress and the Administration have enacted legislation that keeps focused on standards of achievement and gives parents and the public new and meaningful tools for evaluating school performance. An interval of celebration is in order – but most of the tough decisions, and a huge task of implementation, still lie ahead.

Aggregation and Accountability

David Figlio

Introduction[1]

On January 8, 2002, President Bush signed into law the reauthorization of the Elementary and Secondary Education Act. A centerpiece of this education reform involves implementing a system of school accountability. Under the new policy, states must design systems of school report cards based on the fraction of students demonstrating proficiency in reading and mathematics. (States are free to determine how proficiency is measured and defined.) The law requires that states define "adequate yearly progress" in a manner that "includes separate measurable annual objectives for continuous and substantial improvement for…[t]he achievement of economically disadvantaged students; students from major racial and ethnic groups; students with disabilities; and students with limited English proficiency." In other words, states are required to assess schools on the basis of the progress of disaggregated groups of students.

This memo has several objectives. First, I describe the rationale for disaggregating the data by groups. I continue by outlining several of the key potential problems associated with disaggregation, and propose solutions that reduce the pitfalls. To illustrate both the rationale for and pitfalls associated with disaggregation of data, I employ detailed individual-level data covering the academic years 1995-96 through 1999-2000 for every student in two major Florida school districts that must remain unidentified for the purposes of this analysis.

---

[1] Two recent papers raise similar points to those mentioned in this policy memo, and are recommended to the interested reader. The interested reader should consult Thomas Kane, Douglas Staiger and Jeffrey Geppert, "Assessing the Definition of 'Adequate Yearly Progress' in the House and Senate Education Bills," UCLA working paper, July 2001, and David Figlio and Marianne Page, "Can School Choice and School Accountability Successfully Coexist?" forthcoming in *The Economics of School Choice*, C. Hoxby, ed., University of Chicago Press.

Why disaggregate?

There are several arguments, both positive and normative, for why disaggregation of data is warranted. One normative argument centers on fairness; some students, or groups of students, may face schools of different quality. Focusing on the progress of different types of students, the argument goes, may help to ensure that all students are well-treated by the educational system. Indeed, this sentiment is echoed in the alternative name of the ESEA reauthorization, "No Child Left Behind." A related argument is that schools faced with the challenge of improving performance—and with sanctions threatened in case of non-improvement--may seek to help certain groups at the expense of others. Setting performance goals for each population subgroup reduces the ability of schools to "game the system" in this manner.

This paper, however, concerns the more positive arguments for disaggregation of students. The principal positive argument for disaggregation is that different groups of students tend to perform at different levels. Expecting the same levels of school performance without regard to student background and attributes may lead to evaluating schools more on the basis of the composition of the student body than on the basis of any reasonable measure of the school's contribution.

A cursory glance at the data used in this memo makes this point clear. Figure 1 breaks down the variation in 1999-2000 mathematics test scores (adjusted for grade level) into three parts: the fraction of the variation explained by **between-school** differences, the fraction of the variation explained by differences **within a school** in the subgroups identified in the ESEA reauthorization law (major racial/ethnic groups, economically disadvantaged students, limited English proficiency students, and students with disabilities), and the fraction of the variation not explained by either explanation. Put differently, this figure shows the typical range of test scores within each racial (or socioeconomic or other) group in a school, the differences in the typical test scores across these groups within a given school, and the differences in average scores across schools.

The entire pie shown in this figure represents the full range of test scores observed in the data. This pie is, in turn, divided into three slices. One slice, marked "between schools," can be interpreted as the fraction of the total range in observed test scores taken by the range in observed school-average test scores. The second small slice, marked "subgroups within schools," can be interpreted as the fraction of the total range in observed test scores taken by the range in subgroup-average (e.g., comparing average black scores within a school to average LEP scores within a school, and so on) test scores within a school. The remaining slice of the pie reflects the typical share of the total range in test scores observed within subgroups within a given school.

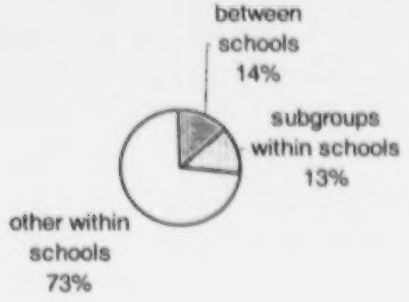


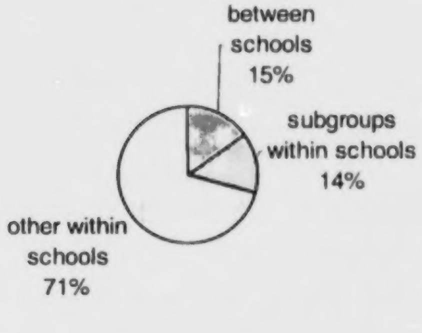
Figure 1: Decomposing the variation in math scores, 1999-2000

As the figure makes clear, the vast majority of the variation in math test scores occurs within schools but is not associated with the general categories mentioned in the law. In other words, there is a great degree of variability in test scores, within a school, in any given subgroup (e.g., among black students in a school.) But the variation that **can** be explained yields informative lessons as well. The within-school differences in mathematics test scores across racial/ethnic/other categories are as great as the between-school differences in mathematics test scores attributable **to any reason.** Given the degree to which schools tend to be racially, ethnically, or economically identifiable, it is reasonable to expect that a substantial fraction of the between-school contribution to variance is actually due to between-school differences in demographic composition. But the striking point from this first look at the data remains that the vast majority in the differences in test scores are not explained by the variables that disaggregation is going to emphasize.

As a further illustration of this point, one can look at the distribution of student test scores within a single school. In one randomly-selected school (which I cannot identify, as mentioned above) with approximately equal fractions black, Hispanic and white, one observes that each of the three racial/ethnic groups are represented among both the top five and bottom five percent of the student body. Put differently, all three racial/ethnic groups have representatives at all performance levels in the school (creating large within-group variation in test scores) even if one group averages higher performance levels than another. This pattern is by no means unique to this particular school, but rather is prevalent in many schools.

Figure 2 repeats the same exercise with regard to reading test scores. It is apparent that the very same patterns present with mathematics are present with reading as well: over two-thirds of the variation in reading performance occur within a school and are not explainable by racial, ethnic, economic, disability, or English proficiency differences. The remaining variation is approximately equally explainable by across-school differences along any lines (including the subgroups of interest) and by within-school differences in the subgroups of interest. (Given that the outcomes of this analysis are qualitatively identical for mathematics and reading, to avoid tedious repetition I will focus exclusively on mathematics for the remainder of this memo. All results described below hold true for both subjects.)

Figure 2: Decomposing the variation in reading scores, 1999-2000



The fact that the majority of the observed test score variation occurs within subgroups within a school **does not** imply that the test score differences across subgroups are unsubstantial. Indeed, there exist dramatic differences in aggregate performance across racial, ethnic, economic, and English proficiency lines that are illustrated powerfully when converting test scores to a standard of proficiency—the requirement of the ESEA reauthorization. Figure 3 shows the fraction of students in a variety of subgroups attaining each of three proficiency standards in 1999-2000—a high standard (one that only twenty percent of students in 1995-96 attained), a moderate standard (attained by forty percent of 1995-96 students), and a relatively low standard (attained by sixty percent of 1995-96 students.) I define standards in terms of fraction attaining in 1995-96 to allow for the possibility of regular, ongoing improvement (or decay) in student outcomes over time while maintaining a constant standard. One observes that some subgroups attain certain performance standards at a rate of two or more times that of other groups. For instance, Asian students are twice as likely as Black students to meet the relatively low proficiency standard in mathematics, and are nearly five times as likely to meet the high standard of proficiency. That said, the fact that the majority of the variation in test scores is not explained by the variables over which disaggregation is to take place suggests that Congress may in fact be pushing states to over-emphasize the disaggregation by "favored" groups as if it were the main reason for test score differences, while the data presented herein suggest that this is not the case.

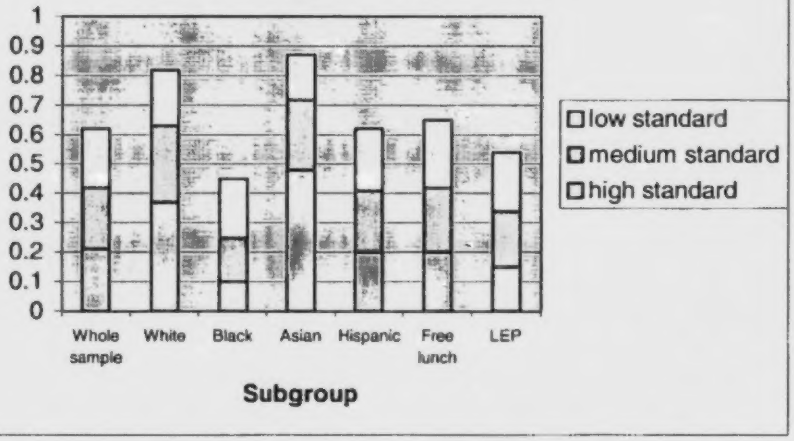


Figure 3: Fraction of students attaining various mathematics standards (standards defined by author)

For the purposes of this memo, there are two major lessons to learn from this glance at the data. First, failure to account for systematic diversity in test performance across subgroups is likely to lead to attributing to schools the test score distinctions due to subgroup differences. Taking these differences into account should lead to an evaluation of schools that more closely measures the actual contribution of schools to student outcomes. On the other hand, if certain groups of students systematically are assigned to the worst teachers within a school (or across schools) then we should not want to control for subgroup differences, as differences in teacher and school quality and adequacy of resources might explain some of the differences being "controlled for."

The second principal lesson to take away from this variance decomposition, however, is that it places a sobering perspective on the use of test scores to evaluate schools. The fact that such a large fraction of the cross-sectional test score variation is unexplainable implies that any aggregated measure of test scores will be measured with considerable error. (Here, I use "error" to mean one of two things: One source of "error" is the presence of systematic differences across groups based on substantive factors unobserved to the researcher. One example might be teacher quality. The other source of "error" is the presence of idiosyncratic positive or negative "shocks" to a student's test score. One example of this type of error might be an unusually poor testing environment on a given day—perhaps road construction taking place outside a third grade class's window.) As measurement error tends to increase with smaller samples, this error may be exacerbated when students are disaggregated into subgroups. Of fundamental import, therefore, when designing a system of evaluating schools is recognizing the trade-off between attempting to more appropriately capture a school's contribution to student outcomes and the further introduction of measurement error that could subvert the accurate assessment of schools. This memo describes this tradeoff in detail below.

The more disaggregated, the better?

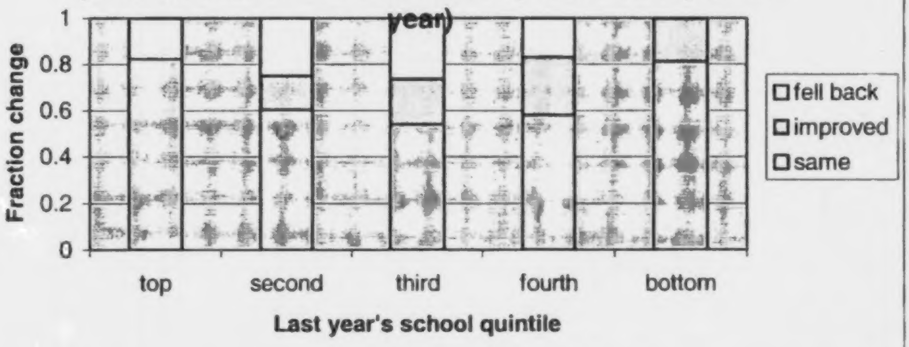
Given the aforementioned reasons for disaggregating, one might be tempted to want to disaggregate data as finely as possible. After all, the argument follows, it makes sense that if measuring progress separately by race/ethnicity and free lunch status better captures a school's contribution to student outcomes, then measuring progress separately for each **interaction** of race/ethnicity and free lunch status would do an even better job (e.g., looking at how students receiving free lunch perform, broken down by race.) Of course, the natural though implausible limit to unimpeded disaggregation would be for each student to comprise his or her own subgroup.

There are, however, legal, practical and methodological reasons for restricting the degree to which student outcomes are disaggregated. Federal law imposes strong restrictions on the publication of individually identifiable student data, and the ESEA reauthorization explicitly rules out disaggregation when students can be individually identified. In the end, what exactly constitutes identifiability would become a judgment call. The practical reason for limiting the degree of disaggregation is that it could become cumbersome to track and interpret the progress of dozens of groups. One challenging issue here involves the definition of "major" ethnic groups. Should all Hispanic groups be lumped together? Should Asian immigrants be treated the same as Asian-Americans for the purposes of disaggregation?

But both of these reasons for restricting the degree of disaggregation pale in importance to the methodological reasons, as for the most part, these methodological concerns would call for a stop to disaggregation well before severe practical or identifiability issues would arise. The principal methodological issue is one of reliability. As mentioned above, most of the variation in student test scores cannot be explained by the observable factors identified in the law. From a statistical standpoint, the remaining variation—be it caused by substantive factors such as within-school differences in teacher quality or by true "noise" like the construction worker's jackhammer—can be thought of as "random." When scores are aggregated over large groups, this randomness will tend to cancel out, and the ensuing aggregates will be rather reliable indicators of "true" performance. But as group size falls, the more likely it is that the "good" or "bad" days of small groups of students could affect aggregate student outcomes, and the resultant indicators would be less reliable.

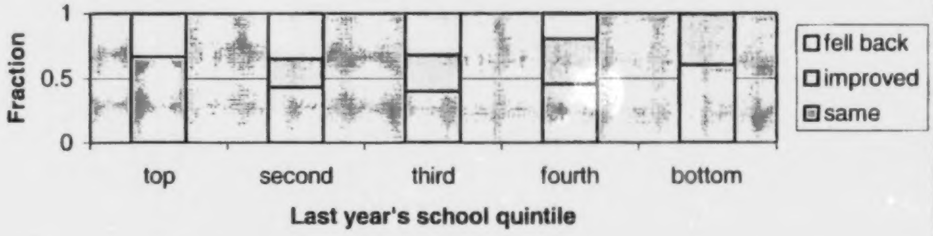
To get a handle on this potential problem, Figure 4 presents evidence on the stability of relative school rankings in mathematics from one year to the next. Here, I divide the schools into quintiles based on the fraction of students meeting a given proficiency standard. (Throughout the memo, I will focus on the high standard of proficiency, as it turns out that the results presented herein are qualitatively very close across all three proficiency measures described above.) As can be seen, about two-thirds of schools remained in the same performance quintile from one year to the next, with the remainder split evenly between improving quintiles and falling back quintiles. While not shown in the figure, only two percent of schools improved or fell back by two or more quintiles. Of course, school improvement and degradation are dynamic processes, and it is reasonable to expect that some schools would substantially improve relative to other schools from one year to the next.

Figure 4: Stability of relative school math rankings (schools that improved to a higher quintile, fell to a lower quintile, or remained in the same quintile from year to year)



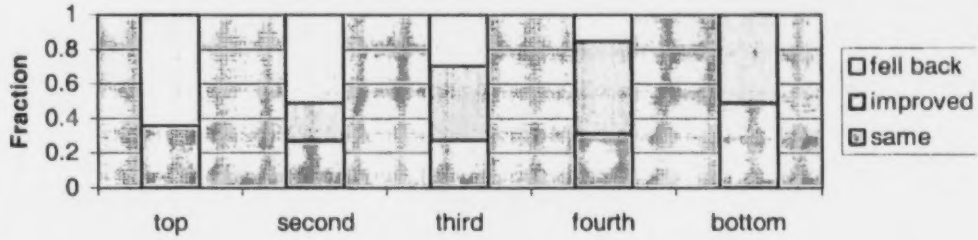
The usefulness of Figure 4 is that it provides us with the opportunity to see how much more volatile relative school rankings can be when a smaller aggregation is employed. Figure 5 presents an analogous exercise in which I explore the stability of school rankings when rankings are based on the performance of black students. Here, unlike in the case of overall school performance, only half of schools remained in the same performance quintile from one year to the next. For instance, in the middle quintile, only about 40 percent of schools remained in the middle quintile in the next year, with the remainder approximately evenly split between moving to a better quintile and falling to a lower quintile. Moreover, fully twelve percent of schools either improved or fell back by two or more quintiles in the rankings. While, as before, some of these transitions may reflect true changes in school performance, the fact that rankings based on black students' performance are considerably less stable than are those based on overall performance illustrates how even a move from the whole school to a large subgroup within the school could lead to measures of school performance that change much more dramatically from one year to the next.

Figure 5: Stability of rankings based on black performance (schools that improved to a higher quintile, fell to a lower quintile, or remained in the same quintile from year to year)



Moreover, it turns out that among the subgroups, the quintile rankings based on black student performance tend to be among the **most stable**. More representative are the quintile rankings for free or reduced-price lunch-eligible students, whose year-to-year transitions are presented in Figure 6. When rankings are based on the economically disadvantaged student body, only one-third of schools remain in the same performance quintile from one year to the next. Remarkably, when ranked based on the performance of free or reduced-price lunch eligible students, more schools changed by two or more quintiles from one year to the next (35 percent) than remained in the same quintile (33 percent). Given that the population of economically disadvantaged students is by no means a very small group (especially in Title I schools, which are arguably the main focus of ESEA,) with 60 percent of the students in the two districts eligible for subsidized lunches, this suggests that rankings based on even large subgroups can be quite unstable over time. This indicates that large year-to-year changes in school rankings could be a substantial problem plaguing any attempt to disaggregate data into student subgroups for the purposes of evaluating schools. Of course, it is possible that these results are driven by low-performing schools focusing additional resources on minority and economically disadvantaged students, but this seems unlikely to have generated one-year changes of the magnitudes presented herein.

Figure 6: Stability of rankings based on free lunch students' performance (schools that improved to a higher quintile, fell to a lower quintile, or remained in the same quintile from year to year)



The issue of persistent improvement is, of course, a central implementation issue for the ESEA reauthorization. Schools are expected to make continued progress toward full proficiency levels over a 12-year period. However, the same errors in measurement alluded to above are only magnified when multiple years of improvement are measured. Figure 7 employs three successive years of data to make this point. In three-quarters of schools in these two districts, a school improved its fraction of students meeting the high proficiency standard in at least one of two possible time windows. But in 69 percent of these cases, a school that improved in one instance fell back in another. This is surely due to measurement problems: Schools facing an unusually "bad draw" (i.e., a large number of students who, for some idiosyncratic reason, did very poorly in one year) one year tended to bounce back in the next, while schools facing an unusually "good draw" (the reverse of "bad draw") one year tended to revert to the mean in the next year. This illustration highlights a central challenge in ESEA implementation; if measurement error makes it difficult for schools to persistently improve their average proficiency, this suggests that compliance with the law will be very difficult to achieve if measured improvement is based on year-to-year changes. This challenge will be further increased if the standards for proficiency also increase over time, as is mandated.

Figure 7: Successive changes in fraction of students meeting proficiency standard from one year to the next

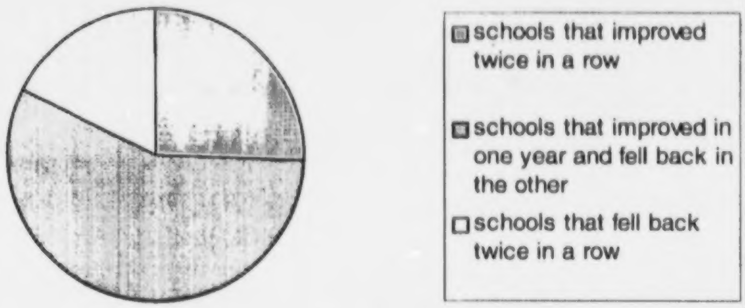
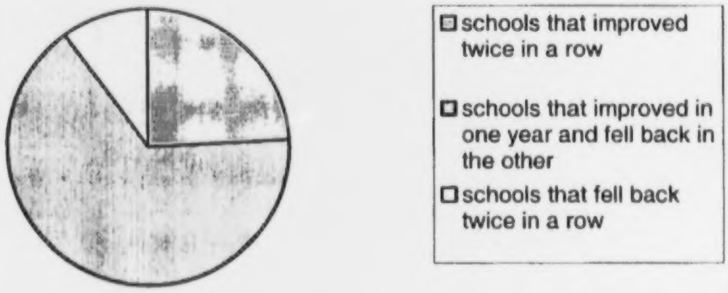


Figure 8 shows that this pattern is unchanged if one restricts improvements and fallbacks to substantial changes (here, measured as changes of two or more percentage points.) Even when restricting the analysis to schools that had two successive substantial changes, we observe that the changes are not persistent—schools that had a substantial improvement in one year were very likely to have a substantial fallback the next, and vice versa.

Figure 8: Successive SUBSTANTIAL changes (+/- 2 percentage points) in fraction of students meeting proficiency standard



As Figures 9 and 10 demonstrate, this pattern is common across all subgroups. The middle segments indicate that in every case, at least half, and usually considerably more, of schools experienced either an improvement in one year and fallback in the next year, or vice versa, when measuring improvements based on the performance of students in the subgroup in question. For instance, when measuring school improvement based on white students, about 60 percent of schools improved in one year and fell back in the other, with the rest about evenly split between improving in both years and declining in both years. When measuring school improvement based on free lunch-eligible students, about 65 percent of schools improved in one year and fell back in the other, while just over 20 percent improved in both successive years and about 15 percent declined in both successive years.

Figure 9: Successive changes in fraction of students meeting standard (ANY improvement)

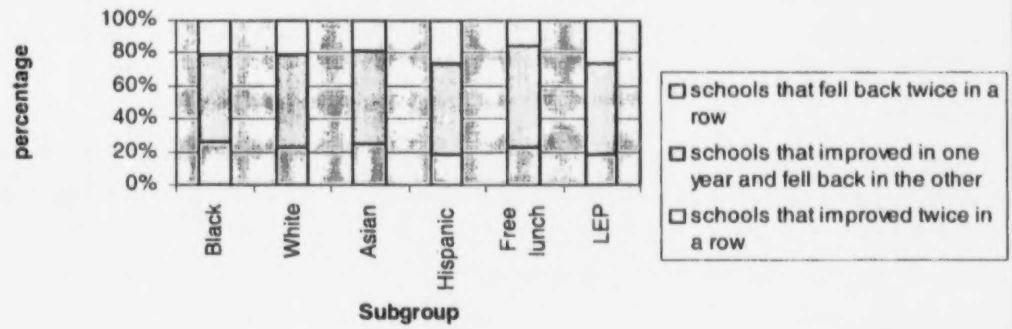
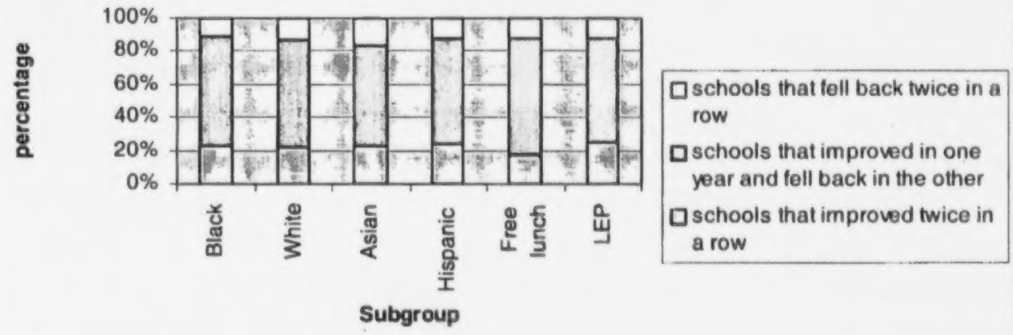
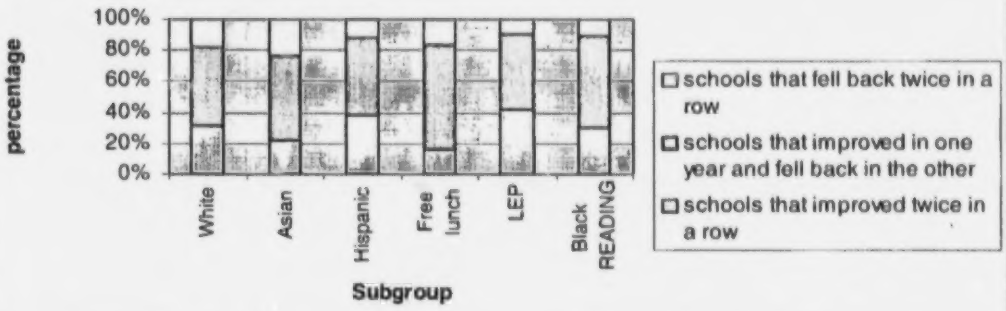


Figure 10: Successive substantial changes in fraction of students meeting standard



It would be more reasonable to trust that the year-to-year changes in the fraction meeting a standard were reflective of true school improvement or degradation if different subgroups were to rise or fall at the same time. To gauge the likelihood that this will occur, one can investigate how performance in **other** subgroups changed over time in schools with a given persistent change in the performance of **one given** subgroup. Figure 11 provides on such illustration of this type of exercise. It illustrates the incidence of persistent improvement and fallback, as well as mixed year-to-year changes for a number of subgroups **for the set of schools with persistent black improvement in math over a three year window**. That is, Figure 11 looks solely at a very select set of schools—the schools that managed to improve their black students' proficiency rates in successive years—and investigates whether these same schools were able to improve the proficiency rates of other students in the same successive years.

Figure 11: Successive changes in fraction of students meeting standard (ANY improvement) -- schools with two years of black math improvement



As Figure 11 makes clear, persistent improvements by one subgroup (black students) do not imply that other subgroups will persistently improve at the same time. Among these schools, as few as 16 percent (in the case of free lunch eligible students) to as many as 42 percent (in the case of limited English proficient students) improved their fraction mathematics proficient in both years in which black proficiency improved. Therefore, the correlation between persistent black improvement and persistent improvement by other subgroups in the same school is rather weak. Moreover, the correlation between persistent mathematics improvement and persistent reading improvement by the same subgroup—black students—is not particularly strong. Only thirty percent of schools with consistent black mathematics proficiency improvements also experienced consistent black reading proficiency improvements. This evidence, coupled with that presented earlier in this memo, strongly suggests that measurement errors might seriously impede the evaluation of schools based on year-to-year changes in proficiency fractions, and that it may be nearly impossible for a school to experience persistent improvements across a wide variety of subgroups. Moreover, the two school districts I am studying tend to have large schools. In much of the country, where schools are smaller, one might reasonably expect the measurement error problems to be even more severe.

Can anything be done?

The previous discussion paints a rather bleak picture of our ability to assess schools' year-to-year improvements. Even in a situation where only school-level aggregates are used to evaluate schools, there exists the very real possibility of serious measurement problems, and as mentioned above, failing to disaggregate reduces the ability to confidently identify observed improvements as school effects. But disaggregation only exacerbates already serious measurement issues.

One partial solution involves reducing the reliance on year-to-year changes in proficiency. The measurement problems described herein come about largely due to idiosyncratic observations, that is, differences in test scores not attributable to the factors being considered by ESEA. But averaging proficiency levels over several years, a remedy allowed under the ESEA reauthorization, can help to smooth out these idiosyncrasies. Figure 12 shows what happens in the data as one increases the time period over which school mathematics proficiency levels are aggregated. We observe that the fraction of "unstable" schools (that is, schools that appear to improve in one year and fall back the next, or vice versa) declines from 57 percent when no moving averages are employed to 33 percent when a three-year moving average is employed. Note that, as a longer time horizon occurs in these data, schools are more likely to have shown persistent improvement. This is not a necessary consequence of employing a moving average; instead, it comes about because this was a period of general secular improvement in standardized test scores in these districts, and the moving average technique brings in more historical data. Similar improvements in data stability occur in the subgroup data, but are omitted from this policy memo due to space constraints.

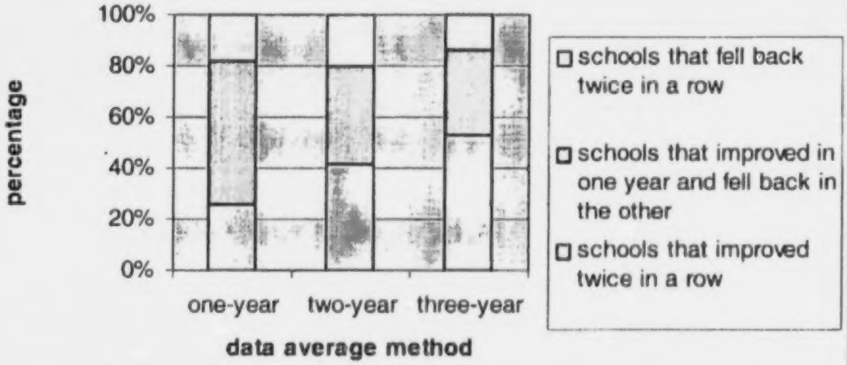
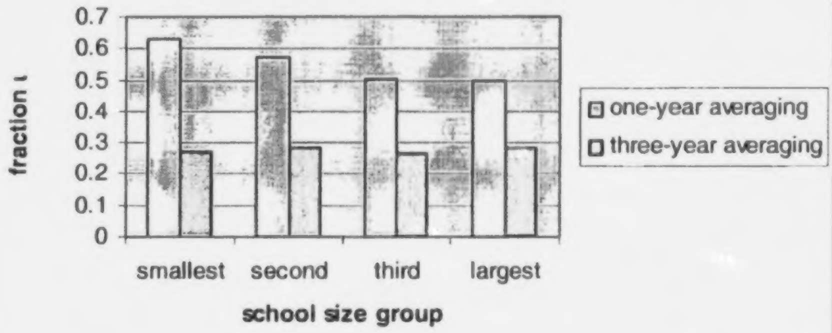


Figure 12: data averaging and stability of observed improvements

In summary, measurement errors associated with "noisy" data can be overcome to a substantial (though far from complete) degree by averaging data over a several-year window. As Figure 13 illustrates, this stability benefit appears to be greatest in smaller schools. In the smallest fourth of schools, going from one-year averaging to three-year averaging reduces the number of schools with "unstable" proficiency changes from 63 percent to 27 percent. In contrast, in the largest fourth of schools, while the three-year averaging yields similar numbers of "unstable" schools (28 percent) to that found in the small-school group, "one-year averaging" (that is, looking solely at year-to-year changes) appears to yield somewhat more stability than found in the small schools. However, even in the largest schools, averaging over multiple years substantially reduces the likely measurement error. Because, as noted, the districts I am studying tend to have larger schools than much of the rest of the country (as well as the rest of Florida) the stability benefits of multi-year averaging might be even greater in other places. Unfortunately, I do not have the data to directly test this hypothesis.

Figure 13: Fraction of schools with unstable proficiency changes--by size class



Even with multi-year averaging, it may still be too demanding a requirement to expect every subgroup to improve in every year, especially in schools with rather small sizes of certain subgroups. I would propose to require schools to improve from year-to-year (using a moving average) for the school population as a whole, but that only some subset of the subgroups need improve in any given year (provided that each subgroup experiences improvement in some fraction of years over a reasonable time window.) A "safe harbor" provision in the ESEA text may create the legal basis for this suggestion. Given the degree of measurement problems associated with even broad subgroup definitions, I would not advocate disaggregating beyond the broad definitions outlined in the law. (With these broad definitions, it is difficult to imagine that confidentiality/identifiability issues will result, except in rare circumstances, which can be dealt with on a case-by-case basis to protect against individuals being accidentally identified.) While the use of moving averages will necessitate a more deliberate introduction of school ratings in many states, it should pay dividends in terms of more stable and believable representations of the contribution of schools to student proficiency.

The fact remains, however, that even with three-year moving averaging, many schools—dozens in any large school district—will have erratic patterns of performance that **may** have little to do with changes in school quality. Strict requirements of consistent growth in proficiency rates from year to year, even with multi-year averaging and even when measured at the whole-school level, may still be too great, and may "punish" some schools even as others avoid interventions that might be warranted.

Comments

Michael D. Casserly

Thank you for the invitation to participate in this important discussion today.

I would like to start this morning by congratulating the Bush Administration and Congress on the development and passage of the "Leave No Child Behind" legislation.

Our organization—and its members—have pledged to work tirelessly to make sure that H.R.1 works as intended—that it improves the performance of all our urban kids.

Our superintendents and board presidents will be coming to town next week to meet with the administration to discuss how to begin translating the bill's promise into reality. Members of the audience may not know that the Council was the only national education group to support the bill.

The bill has a number of features that are particularly important to us and were well articulated in the papers for this conference—particularly Lisa's. One, the bill returns to the original intent of Title I, which was to raise the achievement for our lowest performing kids. Two, it targets scarce federal dollars on the communities most in need. Three, it authorizes the regular assessment of student progress—something urban schools supported enthusiastically. Finally, the measure has a strong accountability system that we also backed.

This is not to say that the bill will be easy to implement. It will not be. We expect to have a great deal of difficulty executing the bill's AYP provisions.

Lisa's paper was correct in stating that few places—state or local—have an accountability system consistent what Congress wants.

Second, we expect to have a hard time implementing the supplemental services portion of the bill.

We do not know what the level of demand will be because we do not know yet where the states will set the initial bars.

Our folks are trying to budget for this right now and are having a very difficult time of it. The problem is compounded by the cuts that states are making in our revenues—at the same time that Congress is increasing them.

Still, we think that a good number of parents will want to keep their kids in our afterschool programs rather than sticking them on another bus.

Third, we expect to have serious problems with the language on fully certified teachers. I have said this before in other settings but let me repeat: we are not passing over certified teachers in order to hire the uncertified one.

We are simply having a tough time recruiting and retaining teachers of all stripes.

Fourth, I think we will have a hard time with the data requirements. We collect almost everything that is asked for—but not in the form Congress wants it or with the facility that the law requires. Our MIS systems are not that nimble.

Finally, our biggest challenge will be getting our instructional programs to do what this legislation envisions.

I was particularly impressed by Lisa Graham Keegan's paper. The only item I was in disagreement with was the assertion that the bill's "safe harbor" provision would solve the problem about identifying too many schools as failing. I think the safe harbor provision will have little effect beyond the margins.

You may also want to revisit the language requiring districts to spend 20% of their Title I funds on supplemental services. The bill doesn't quite say what your paper does.

Otherwise, I was in complete agreement with your points on focusing AYP on progress not compliance; on the importance of disaggregating results; on the inherent flexibility in the law; on the importance of NAEP; and on futility of arguing about the relative merits of criterion and norm-referenced testing. It was extremely well-done.

Your call for humility and flexibility among state leaders in the implementation of the bill was well stated and on point.

We will be urging the department to adopt a "rational basis" test while implementing the bill—rather than regulating on each and every clause.

I was not as taken, however, with the paper prepared by the New American Schools.

I agree with you that the bill does not really measure progress—it measures proficiency. But, Congress heard our arguments on this point and clearly rejected them—including the use of longitudinal data to demonstrate growth on AYP.

You also make several points in your paper about charter schools and why they should be exempt from one provision or another or considered for grace periods.

Your line of argument strikes me as exactly the kind of excuse-making that we used to do and that the public rejected.

You won't get any sympathy from us.

I think the law is very clear that the accountability provisions apply to charter schools like they apply to us—although there is some flexibility about who does the monitoring. There is no flexibility on the point that they apply to public charters

Finally, I agree with David Figlio about the technical problems with the disaggregated data. The data are going to fluctuate for all kinds of reasons.

We ran into many of these anomalies last year when we were preparing our report, *Beating the Odds*—which, by the way, we will re-release later this Spring with updated data.

But, we thought it was just better to start releasing imperfect data than to wait until all the technical problems are solved.

As I indicated, this bill will not be easy to implement, but we supported it because it had all the right goals; for all the right reasons; focused on all the right places.

The nation's urban schools will do everything to make it work.

I think that is what the nation wants from us. And that is what we plan to give them.

Thank you.

Implementing Title I Standards, Assessments And Accountability: Lessons From The Past, Challenges For The Future

Michael Cohen

In 1994 President Clinton proposed and Congress passed several pieces of sweeping legislation that, for the first time, forged a state-federal partnership to implement standards-based education reform nationwide. Goals 2000 and the 1994 ESEA reauthorization (the Improving America's Schools Act) created a new framework for the federal role in elementary and secondary education, based on challenging state standards and aligned assessments for *all* students, accountability for results, flexibility in how to achieve them, and increased targeting of federal education resources to high poverty schools. These programs combined the federal government's historic role in providing aid to schools serving our nation's most disadvantaged students with a new effort to ensure that federal education programs and resources supported the implementation of state and local standards-based education reforms.[1]

In 2002 Congress again reauthorized ESEA, enacting major components of President Bush's No Child Left Behind Act. This legislation builds squarely on the foundation laid in 1994, and extends it by providing state and local education officials with greater flexibility in the use of federal resources, significantly tighter school accountability requirements, and greater federal direction over the design of state testing and accountability systems.

Most significantly, it changes the ground rules for accountability, by requiring schools and school systems to bring every child up to state standards within a finite period of time, and to close achievement gaps based on race, ethnicity, language and income. Persistently low performing schools and those that succeed for *some* students but not *all* will be under considerable pressure to find ways of effectively addressing the needs of students being left behind. More forcefully than before, these new provisions take aim at the "tyranny of low expectations" for students from poor and minority backgrounds, and seeks to replace them with a culture of high expectations and adult responsibility.

---

[1] This paper draws heavily on my experience as Assistant Secretary in overseeing the implementation of the 1994 Title I requirements for standards, testing and accountability,

The requirements for states to implement systems of standards, assessments and accountability have been the central feature of federal elementary and secondary programs since 1994. They have also been the most politically difficult to craft and implement. Their successful implementation depends upon the willingness and ability of federal and state officials to negotiate a complex set of technical, political, legal and organizational challenges, and a good deal of luck.

The actions taken by state and federal officials to address these requirements will determine whether they provide the right pressure to drive needed changes in state and local policy practice and resource allocation, prove to be unworkable on the ground, or are relegated to the margins of state and local education reforms.[2] The experience of implementing the 1994 requirements can shed some light on the challenges and opportunities and choices facing federal and state officials this time around.

IMPLEMENTING THE 1994 TITLE I REQUIREMENTS: PROGRESS IN THE STATES

In the eight years since Goals 2000 and ESEA were enacted, states have made considerable progress in some areas, and far less in others. More specifically:

Content and Performance Standards

- 49 states had adopted state content standards in the core areas of reading/language arts and math.

- Few states met the 1997 deadline for implementing performance standards in these same subjects and grade levels, though virtually every state is expected to meet this requirement at the same time it completes the implementation of required assessments.

States had relatively little difficulty complying with the requirements to adopt content and performance standards in reading and math. Note however that most states received waivers of the statutory deadline for setting performance standards, because the law envisioned that performance standards would be set prior to the development and implementation of assessments, while states found it necessary to first develop and at least field-test the assessments before setting the cut scores for proficiency levels. Note also that compliance hasn't ensured quality. Periodic reviews of state standards done by Education Week, Achieve, AFT and the Fordham Foundation over the years reveal that state standards vary considerably with respect to rigor, clarity and other dimensions of quality. Further, these rating systems don't always agree with one another, underscoring that as a nation and an education community we are still learning how best to define and implement the standards.

---

[2] Whether they produce the intended results for students will depend *most* heavily on the level and type of day-to-day support, professional development, tools and resources teachers and principals receive from local, state and federal officials, but consideration of that is beyond the scope of this brief paper.

Aligned Assessments for All Students

- As of January 2001, 17 states were on track to meet the 2000-2001 school year deadline set by ESEA for having aligned assessments, that included all students, in reading and math at least once each in the elementary, middle and secondary school grade spans. An additional 14 states had received a waiver of the implementation deadline, but were still clearly on track to meet the requirements with some additional time.

- In contrast, 3 states – California, Wisconsin and West Virginia – were found by the U.S. Department of Education to have been substantially out of compliance with the requirements and not likely to meet them unless forced to do so by the federal government. Subsequently, Alabama was added to this list.

- In meeting the federal requirements, many states were required to change assessment practices they had already adopted. In particular:

    - Nearly all states were required to take additional steps in order to include all students, especially limited English proficient students and students with disabilities, in their assessment system. These steps include ending time-based exclusions for LEP students, and the provision of accommodations for students who need it due to language or disability.

    - Thirty states were required to modify their procedures for reporting school-, district- and state-level achievement data in order to provide disaggregated data on student performance, based on race, ethnicity, gender, family income, disability and migratory status.

In general, states faced much greater difficulty in complying with the requirements for aligned assessments than for establishing standards. Full compliance required significant changes in test design, administration and reporting practices. Meeting these requirements required states to shift from norm-referenced to standards-based assessments, and to end long-standing practices of excluding students with disabilities and limited-English proficient students from the state testing, reporting and accountability programs. In addition to the considerable technical challenges in meeting these requirements, many states had to respond to legislative and parental demands for norm-referenced achievement data, local desires to continue existing, locally determined testing programs, and concerns among educators about being held accountable for education LEP and students with disabilities to the same standards as all other students.

Identifying and Intervening in Low Performing Schools: Adequate Yearly Progress and School Improvement

- States varied tremendously in key design and performance elements of their accountability systems, and in the number of schools identified each year as "needing improvement".

  - While all but 5 states had set absolute goals for school performance, they varied considerably in the percent of students expected to meet state proficiency standards (about a dozen states expect 90 – 100 percent of students in each school to meet the state's proficient standards, while another ten set a goal of 50 percent of students meeting standards in order for a school's performance to be satisfactory.)
  - Only 14 states set specific timelines for meeting performance goals (on average, ten years, with a range of six to twenty years).
  - States used vastly different methods for defining adequate yearly progress. Some states required schools to meet an absolute performance target, while others required relative improvement each year or reductions in achievement gaps between subgroups of students, and yet others used various combinations of these approaches.

- States also varied tremendously in the proportion of Title I schools identified as "needing improvement". At the low end, Texas identified only 1% and North Carolina identified less than 5%; at the other extreme, Michigan identified 76% and Washington DC identified 80% of Title I schools as needing improvement. In school year 1998-99, approximately 20% of Title I schools were identified as needing improvement, and that number has been increasing annually.
- According to a national survey of principals of Title I schools, only about half of the schools identified as needing improvement received any help (e.g., professional development, technical assistance, additional resources), although both the provisions and the logic of the law required that they do.

LESSONS FROM THE PAST

We have learned some important lessons over the eight years since these two laws were enacted, that can and should inform how the federal government and states proceed with implementation of the new law.

Federal legislation pushes all the states forward – even if they don't all comply with the letter of the law.

In 1993 when the Clinton Administration took office, only a handful of states were developing standards and aligned assessments and preparing to use them as the cornerstone of their education reform strategy. Now, nine years later, every state is organizing its K-12 system around standards-based reform, and there is little debate about the *appropriateness* of this direction, though there is much vigorous debate about the *quality* of the design and implementation from state to state. The point here is simple but important: since the federal government is the junior partner in education, much of its impact occurs because of the overall direction it provides, not solely as a result of the specific strings tied to federal funds. New or significant changes in existing federal programs frame the terms of the debate and policy deliberations that go on in every state and community, mobilize supporters, and create an expectation for action consistent with the new law. Though the standards movement began among the states, there can be little doubt that the combination of Goals 2000 and the 1994 reauthorization of ESEA helped move every state, as well as the broader education, business and policy communities, in that direction.

If it can't be done, it won't be done.

States won't implement requirements that are unworkable or meet deadlines that can't be met, no matter what the law says.

The 1994 law required states to establish content and performance standards in reading and math by the 1997-98 school year, and final assessments aligned with the standards by the 2000 – 2001 school year. While most states met the deadline for content standards, almost none met the deadline for performance standards, and some states still haven't. Notwithstanding the statutory timeline, states found it almost impossible to develop, define and describe performance standards in the absence of the assessments that made them real and concrete. To accommodate this situation, the Education Department agreed to provide waivers for the implementation deadline for performance standards.

That decision was right, but not without consequence. In the context of other factors discussed below, freely granted waivers contributed to a belief that the Department would not in fact enforce any of the requirements for standards, assessments and accountability, and may therefore have undermined state compliance over time.

I'm convinced that the timeline for implementing final assessments in the 1994 law – by the 2000-2001 school year – was workable, if a state started development work soon after the law was enacted. However, many states delayed the development of assessments for several years. Once that delay occurred, and once a state began good faith efforts to develop the assessments, there was little either the state or the federal government could do to speed up the process. Developing and field-testing items and performance tasks, developing scoring procedures, conducting validity and reliability studies, all take time. And as the school year 2000-2001 implementation deadline approached, there were a number of states working in good faith to complete development work, but which would clearly not meet the deadline. No threatened or real sanctions could speed the development and implementation timeline at that point. The Department's response therefore was to waive the statutory deadline and hold the state to sticking with the timetable it was already on.

Some view these decisions as evidence of *lax* enforcement by the Department, though I see it as evidence of the *limits* of enforcement.

These limits will become apparent again soon, because some states will not be able to meet the deadlines in the new law. For example, the results from tests administered this spring are to form the baseline for defining adequate yearly progress. However, those states that still do not have final assessments in place, or that plan necessary changes in their current assessments, will be unable to use this year's results for that purpose, since they will be using a different test in the next year or so.

If we don't know how to do it well, it will probably be done poorly, if at all.

Clearly the most disappointing aspect of the implementation of the 1994 requirements is the fact that the states are literally all over the map with regard to adequate yearly progress and school improvement. States vary widely with respect to their content and performance standards, assessments, time frame for expecting all students to meet proficiency standards, rate of progress and/or the basis of comparison necessary to be considered adequate, and the proportion of Title I schools identified as "needing improvement". And large numbers of schools that are identified as "needing improvement" by *any* definition of adequate yearly progress receive little or no help in order to improve. And each year, the number of schools identified grows. The major problem here is not that too few schools are identified as low performing, but that once identified, too few of them get the help they need to improve.

I am convinced that *one* of the reasons for this is that states, and the education community overall, do not have a clear, research-based idea of how to effectively set performance or progress targets for individual schools or school districts, and there is much confusion about the technical requirements and strengths of various approaches. Similarly, we still know relatively little about how to organize and implement school intervention strategies on a sufficiently large scale and for a sufficiently sustained period of time, in order to turn around high poverty, low performing schools.

As states worked to define adequate yearly progress, they had little or no experience or research to clarify the rate of progress they could reasonably or even ideally expect any given school to make. How could they – for the standards against which schools would be measured had just been set and, in most cases, had not been implemented for more than a year? The experience of other states, a couple of years ahead at most, was difficult to draw upon, for there is no way to equate the performance standards from one state to the next.

As many of the other papers prepared for this conference demonstrate, there is considerable confusion about the technical feasibility of effectively implementing some of the new adequate yearly progress requirements. To the extent that there are design problems built into the statutory requirements, or insufficient technical capacity in states to implement them well, we can expect that there will be minimal or inconsistent compliance with the requirements – a situation that can't be altered by aggressive enforcement efforts by the U.S. Department of Education.

To help states meet the new requirements, the Education Department must support the R&D, evaluation, policy analysis and networking among states necessary to help state education officials learn how to design effective and workable accountability systems.

Congress rarely provides sufficient resources for this purpose. In addition, by its nature, such research is slow in coming – it must follow, not precede implementation, and will therefore almost always be too late for the initial design decisions states make. The knowledge base about turning around low performing schools once they are identified is also limited. To be sure, there is an important knowledge base about the characteristics of high poverty, high performing schools as well as a growing number of examples of low performing schools that have successfully been turned around. But only recently have we even begun to develop a robust set of research findings and practical experience that would help state and local education officials organize a statewide capacity for intervening in a significant number of low performing schools.

At the same time, we've learned with greater certainty that many of the problems plaguing persistently low performing schools can be traced to the relatively large numbers of, and high turnover rates among, the poorly prepared and inexperienced educators who staff and lead them. While there is no doubt that states and local school districts can and must do much more to attract and retain highly qualified teachers and principals, addressing this issue is generally well beyond the capacity of school intervention teams organized by state and local education agencies.

Each state marches to the beat of its own drummer – and sometimes, more than one drummer.

Consider the following examples:

- ☐ Assessment systems in California, Wisconsin, West Virginia and Alabama were found by the Department of Education to be substantially out of compliance with Title I requirements, and the states were told they must enter compliance agreements in order to remain eligible for Title I funding. In each of these cases a substantial portion of the compliance problem can be traced to decisions by the legislature, generally with the support of the governor, to mandate norm-referenced tests that were not aligned with state standards. The conversations I had with the chief state school officers in those states made clear that the legislature acted without much knowledge of or attention to the Title I requirements.

- ☐ This past September the National Conference of State Legislatures sent House and Senate conferees a letter expressing its concern that "...the testing requirements at the heart of both [House and Senate] bills is an egregious example of a top-down, one-size-fits-all federal reform." The letter went on to express similar views about a number of related provisions. Though the heated political debates about "federal intrusion" into state education matters that reached a fever pitch around Goals 2000 has abated considerably in the past several years, there are still serious concerns among state policymakers nationwide about the extent to which the federal government should attempt to dictate the specifics of state testing and accountability policies. This letter is probably not the last word from state legislatures on this subject.

- In a paper delivered at a Brookings Institution conference last Spring, Paul Hill and Robin Lake described the failure of the Washington State legislature to pass a long-awaited school accountability bill. Both the paper and my own conversations with state officials and informed observers underscore the widely held view that, as a result of the legislative stalemate, there is no system for school accountability in the state. Apparently no one involved in this debate (pretty much the entire education, business and state policy communities) thought that the combination of the 1994 Title I requirements, along funds from the CSRD Program, the Reading Excellence Act and a new Title I accountability fund established by Congress in FY 2000 in any way constituted the basis for a state accountability system. As far as I can tell, the existence of the federal requirements and funding didn't even enter the debate.

- In the January 16, 2002 edition of *Education Week*, the front page headline states "States Gear Up for New Federal Law" while a story on page 16 reports that "Michigan Chief Sees School Ratings, Sanctions in Future." According to this and previous *Education Week* accounts, the new chief state school officer in Michigan replaced the never-quite-implemented, test-focused accountability system with a new school grading system that relies on a broader set of indicators of school quality, including family involvement, quality of professional development, attendance and dropout rates, among others. Under the chief's proposal, schools would receive their first "grades" in 2003, but sanctions could not be applied until 2005.

  Almost none of the accountability features reported in this story appear to comport with the requirements in No Child Left Behind signed into law the previous week, and evident to most observers since last Summer. NCLB permits multiple measures, as long as those added to the state assessment program don't reduce or change the schools identified for improvement. Interventions in low performing schools must begin immediately, at least for those already identified under the 1994 provisions. Michigan has already identified some 80% of its Title I schools as needing improvement, so there should be widespread interventions occurring now, not delayed until 2005. In brief, it appears that, at least in Michigan, the state's accountability system and the Title I accountability system operate in parallel universes.

These examples highlight several important lessons. First, in many states the governor and the legislature, not the chief state school officer and the state education agency, are in charge of testing and accountability policy. In general, legislators and governors don't pay any attention to Title I requirements, and may not even be aware of their existence. While the relevant provisions in NCLB generated a fair amount of media attention, few state policymakers will give it much thought six months to a year from now – let alone by 2005 and beyond when new testing requirements must be implemented. The odds are pretty high that governors and legislatures in most states will continue to think they have a free hand on these issues. This means that if the Education Department wants to ensure state compliance with these requirements, it must launch a sustained communication strategy targeted to legislatures and governors. The Secretary must explain to them that, from now on, he is their partner when it comes to testing and accountability policies.

Second, within state education agencies there is all too often a wall between federal program coordinators and those responsible for the overall development and implementation of state education policies. For example, it was clear when we reviewed state assessments, in some states the state testing director had not worked closely with the Title I, Special Education and Bilingual Education program staff, which is necessary in order for the state to have a clear and coherent testing, reporting and accountability system that meets federal requirements. While the Education Department has made vigorous efforts in recent years to communicate with all relevant offices in state education agencies regarding these requirements, it is ultimately up to chief state school officers to ensure that they create the sustained internal communications necessary to support effective implementation of the new requirements.

No one believes the Education Department will really enforce Title I requirements.

Consider the following examples:

- When I was preparing to become Assistant Secretary for Elementary and Secondary Education, I began to talk with others inside and outside the Administration about the importance of vigorously enforcing the Title I assessment requirements that were due to be implemented in the immediate future. A number of trusted friends -- experienced political appointees, career staff and old Washington hands -- all told me the same story: When Frank Keppel was Commissioner of Education under LBJ, he attempted to withhold federal education funds from Chicago because it failed to comply with certain desegregation requirements. Upon receipt of formal notification of this from the Department of HEW, Mayor Daley called LBJ directly to complain. Frank Keppel was gone by the next day. And no one has been foolish enough to try anything like that since.

- In late 1999, at my first meeting with the program directors and senior staff in the Office of Elementary and Secondary Education after becoming Assistant Secretary, I announced that my top priority was to ensure that states fully complied with the Title I assessment requirements, and that we would launch a campaign to persuade the states that we would use all the enforcement tools at our disposal to ensure compliance. Most of them looked at me like I was nuts, and a few of them politely indicated that this would be a new direction for the office.

- When Congress was considering the Clinton Administration's ESEA reauthorization proposal in 1999, I told a House committee staffer that I thought an accountability provision they were considering went too far. His response was that they planned to hold firm to their current position, confident that the Senate would pass a more watered down provision they would have to compromise with in conference, and that the Education Department would completely water down whatever emerged in the final bill.

The fact of the matter is that the Education Department does not have a strong track record of compliance monitoring in ESEA programs, and hasn't for decades spanning Administrations of both parties. There is a widespread view that the Department has few effective sanctions to apply, since no one believes that it will ultimately withhold funds from states or local districts. When I became Assistant Secretary I realized that the Title I program, for a number of reasons, had an inconsistent record of compliance monitoring. It lacked both the staff capacity and the clear focus to pay attention to the most important requirements, and send clear and consistent messages to states about the need to meet them.

This longstanding track record was surely compounded by both the intent of Goals 2000 and the politics surrounding its enactment and implementation. Goals 2000 was intended to help states jump-start standards-based reform, while deliberately providing them with a great deal of flexibility in the design of state reform strategies and the use of federal funds. While the Secretary was required to review and approve each state's education reform plan, we were keenly aware that Goals 2000 provided less than one a percent of total state education expenditures, and worked hard to ensure that the peer review of a State's plans recognized the limits this imposed. Many states understood and appreciated this approach, but others saw it as another indication that the Education Department lacked the will for tough-minded compliance monitoring.

The political assault on Goals 2000 in a number of states (e.g., the governors of Virginia, California, Alabama and New Hampshire refused to accept Goals 2000 funds because of the "federal intrusion" and "strings" that came with it[3]) coupled with simultaneous efforts in Congress to abolish both Goals 2000 and the Education Department, also contributed to a widespread view that tough enforcement would be a particularly hazardous course of action for the Department to pursue.

The states are and must remain the "laboratories of American democracy"

For the past 20 years, states have been the driving force in education reform, in large part because they have both the responsibility and the room to find their own solutions to common challenges, and a fairly robust tradition of learning from one another. No Child Left Behind will heighten the attention paid to testing and accountability in every state and district. As states work to address these issues they need the room to take advantage of emerging solutions and opportunities, and to address problems not foreseen in, or perhaps created by, the legislation. In working to secure full implementation and compliance with these new requirements, the Education Department must ensure that it does not become an obstacle to needed progress. At least several areas come to mind as illustrations of where the Department must find ways to ensure that the Title I requirements don't become obstacles to needed state experimentation that can lead to improved practice and better results for students:

---

[3] Ironically, these same governors had little problem accepting Title I funds, though that program had far more specific requirements for state standards and assessments. More ironically, Alabama and California then proceeded to ignore the requirements.

- The Adequate Yearly Progress and School Improvement requirements apply equally to high schools as to elementary schools, yet it is not clear that they make as much sense at the secondary level. For example, states with high stakes high school graduation requirements must find effective ways to intervene in high schools with high failure and/or dropout rates, even if the percentage of students passing the test increasing significantly each year. A school that increases the pass rate from 65% to 75% in a year may be making exceptional progress but it is hardly adequate if a quarter of the students can't meet the graduation requirements. Consequently, many states will need to find different yardsticks for judging the performance of high schools, and more powerful and swift intervention strategies than the graduated series of steps provided for in statute. The Title I requirements should not be a barrier to effective state action in this area.

- Online assessment appears to offer many advantages for states, teachers and students. It holds the promise of immediate results and feedback so that the tests can be used to improve teaching and learning for the students who take them. They can be administered at different points in time, enabling students to take them when they are ready to demonstrate they have met the standards, rather than on a single "one-size-fits-all" testing date. They may be customized for individual students, enabling students to take fewer questions that are better geared to their level of performance, potentially increasing both the efficiency and the diagnostic value of the tests. Yet the Feb. 6, 2002 issue of *Education Week* reports Education Department officials have indicated that Idaho's approach to online testing may not meet Title I requirements. The particulars of the Idaho situation will matter a lot in the final determination, and there may be other ways to use online assessments that clearly fit federal requirements. But in general, the Title I requirements must not become a barrier to the necessary development and experimentation in states.

- The annual testing in grades 3-8 required by NCLB will make it possible for states and districts to use "value-added" approaches to measuring the performance of schools, and identifying as needing improvement those schools that make little contribution to student achievement each year. It isn't clear whether this approach is necessarily superior to the cohort approach NCLB builds in to the definition of adequate yearly progress, but it certainly deserves serious consideration. In any event, it would almost certainly identify a different set of schools as low performing than the prescribed approach, but its not clear it is permissible under the statute. The Title I requirements should not be a barrier to sorting out the most appropriate approaches to identifying low performing schools.

The bottom line here is simple. We don't yet know all we must in order to translate the principles guiding NCLB into the most effective actions. Yet the specific requirements appear to leave some approaches off the table, even if they may turn out to be more promising. While the Education Department has a clear responsibility to ensure that every state complies with the new requirements, it also has a responsibility that to help states find the most effective approaches to meet the overall purpose of improving achievement and closing achievement gaps.

A cautionary note: when it comes to accountability, too many states opt to be in the control group in the laboratories of American democracy. In 1985 the National Governors' Association issued a landmark report, *Time for Results*, in which the governors urged each other to ensure that every state adopt policies for turning around "academically bankrupt" schools and school districts. The governors have returned to this theme of tough-minded accountability virtually every year since, including at three different education summits in the past decade. In 1986 nine states had such policies in place. The 2001 *Quality Counts* report on state education policies shows that now, twenty-seven states have state policies in place for identifying and intervening in low performing schools – and far fewer provide all low performing schools with external assistance and additional resources. By now, every state should, even without federal requirements, yet the number of states that do has been increasing by only about one state per year.

At this pace, it will take another twenty-five years until all states do, way too long in an era in which education is an urgent national priority. The fact that states have constitutional responsibility for education should afford them considerable leeway in determining *how* best to deal with persistently failing schools. It should *not* mean that each state can choose whether or not to turn its back on the situation altogether. Frustration with this slow pace accounts in large part for the strict and specific provisions Congress has now required states to implement.

A balance of flexibility and focused enforcement can work.

Notwithstanding the obstacles identified above, it will be possible for the U.S. Department of Education to secure substantial state compliance with the Title I requirements, with an approach that balances adequate flexibility for states to implement new requirements in ways that fit their approaches and circumstances, and firm insistence to essential, nonnegotiable requirements. This approach should include:

- ☐ Working in partnership with states to find effective and appropriate ways to meet new federal requirements in ways that are most consistent with each state's overall reform strategy and direction. This means, for example, working with states such as Nebraska and Maine to help them continue with their efforts to allow for local assessments aligned with state standards, instead of a uniform statewide test. This should be distinguished from helping states that will want to continue to with a patchwork quilt of state tests in some grades, and varying and unaligned local tests in other grades.

- ☐ Articulating a clear set of priorities for enforcement and compliance monitoring, and communicating about them clearly and consistently to all of the appropriate state officials, and those who inform and influence them. This means that the Secretary and other senior Administration officials must make clear to governors and legislators that there are testing and accountability requirements that each state *must* comply with, without exception.

- Being prepared to use a full range of enforcement strategies – from jawboning to compliance agreements to withholding administrative or program funds if necessary. States must think that *all* of these are on the table. If the Department interprets the statute to limit the enforcement tools solely to withholding some or all of state administrative funds, the Department's ability to secure state compliance will be seriously eroded. I can think of a number of states where the governor and legislator would not view the prospect of reduced funds or staff for the state education department as a serious sanction.

- Building and maintaining a strong monitoring and assistance capacity within the Education Department. At a minimum, this requires:

    - *An implementation team*, led by senior Administration officials and including a capable, experienced and stable team of career staff. While the Administration must provide the leadership and make the final policy calls, this work can't be done unless there is a core group of a half a dozen or more seasoned and technically knowledgeable staff from the program office and the general counsel's office assigned to working with states. They are the ones who must stay in regular communication with state staff, provide guidance and support on a range of specific issues, and ultimately manage the process of reviewing detailed, technical and voluminous submissions from states.

    - *A state-by-state monitoring strategy*, that starts by working with the relevant officials in each state (including the governor and legislature as well as the state education agency) to determine a plan and timeline for closing gaps between current policy and the new requirements. The Education Department must then regularly monitor implementation in each states, and help the state stay on track. The Department should be prepared to help states fit their own approaches with the federal requirements. It should also be prepared, *on a carefully selected basis*, to waive deadlines when a state cannot possibly meet the deadlines or when doing so would result in costly disruptions. Similarly, the Department should be prepared to consider requests for waiving specific requirements if the state has a sound approach to accountability consistent with the Title I principles and purpose, but that does not meet all of the specific requirements.

    - *Reporting requirements and data tools* that will enable the Department to monitor state implementation of key provisions in a timely fashion and with minimal unnecessary burden on states. This means insisting that states provide annual performance reports in a timely fashion – in the past, many states took an additional six to twelve months to provide the Education Department with needed data, despite considerable efforts by Department staff to secure the reports. It also means continuing the Integrated Benchmarking and Performance System (ITBS), a partnership with states to help develop and implement electronic mechanisms for "harvesting' data in states' electronic warehouses. Such a system could provide data on every school state's have identified as needing improvement, including data on teacher qualifications and student achievement gains each year, without the need for paper reports from states and local districts.

- Following through on all of the compliance agreements that were set in motion under the previous Administration. To the extent that these are replaced with waivers, they will be seen in the field as a retreat from significant enforcement – and a sign of the Department's stance in the coming months and years.

CONCLUSION

This paper has focused on the factors that can affect state compliance with the new Title I standards, assessment and accountability requirements. While these are important, even complete compliance will not be enough to bring about necessary gains in student achievement. Translating tougher accountability measures into large scale achievement gains for all students will require substantial investments at the federal, state and local levels to recruit, prepare, and retain talented teachers and principals, to support them with the high quality professional development, curriculum and instructional materials aligned with standards, and tools to support data-based decisions. It will also require substantial investments to give students the opportunities to learn, included smaller classes, modern buildings and 21st century technology, and extended learning opportunities through after-school and summer programs. Attention to compliance must complement, not substitute for, action in these other areas.

What Might Go Wrong with the Accountability Measures of the "No Child Left Behind Act?"

Dan Goldhaber

On January 8, 2002, President Bush signed the reauthorization of the Elementary and Secondary Education Act (also referred to as the "No Child Left Behind Act"). In many ways the passage of this legislation marked a significantly more prominent federal role in education. This is especially true with regard to the accountability provisions, which suggest that the federal government will, for the first time, penalize schools that fail to achieve "adequate yearly progress," as defined by student performance on standardized tests. Rewards and sanctions are, of course, designed to lead to better student outcomes, but incentives that are not properly structured may result in policies and behaviors that are not universally beneficial. In this memorandum, I explore the potential pitfalls associated with this new federal accountability role. In doing so I am not arguing that these worst case scenarios described below are likely, only that it is well worth the time to consider the potential for unanticipated negative consequences so as to try to avoid pitfalls before they occur.

There are, of course, many potential unanticipated negative consequences associated with any accountability system, be it at the local, state, or national level. After providing a general overview of the new federal, state, and local accountability relationship, I will focus on how accountability systems may create unanticipated negative consequences. The hope is that by pointing out the possible pitfalls associated with a federal role in accountability, these pitfalls may be avoided.

Overview of the New Federal Role

The centerpiece of the new federal role in accountability is the requirement that states administer *high-quality* annual academic assessment tests in reading and math for every child in grades three through eight by the 2005-06 school year. (In 2007-08 schools will also be required to administer annual tests in science.)[1] These assessments must be *aligned* with standards, *consistent* with nationally *recognized* professional and technical standards, be *used in a valid and reliable manner*, and *test higher order thinking skills* using multiple measures.

Each state is required to create a system of rewards and sanctions based on whether students from a number of different sub-groups make adequate yearly progress (AYP) towards the state's *proficient* level of academic achievement.[2] AYP must be defined so that in each state all students in each group meet or exceed the state's proficient level of academic achievement "not later than twelve years after the end of the 2001-2002 school year" (2013-14).[3] Schools that fail to demonstrate AYP for two consecutive years are required to provide students with additional public school choices. If schools fail to improve after a third year, parents of students in those schools may use a portion of the school's Title I aid to purchase supplemental educational services, including private tutoring. Schools failing to improve for five consecutive years may be subject to reconstitution.[4] The legislation also requires states to participate in the National Assessment of Educational Progress (NAEP) in reading and math, which means a sample of students from the state will take this national proficiency test in grades 4 and 8. Student performance on the NAEP will be used to *verify* reported performance on the assessments used in each state.

While few argue against "appropriate" accountability measures, debate arises in regards to what is appropriate, and, in the case of the reauthorization of ESEA, the devil is very much in the details, many of which are sketchy and left open for negotiation between states and the Department of Education. For example, the question of what constitutes adequate yearly progress received a great deal of attention.[5] AYP along with the other italicized words and phrases in the preceding two paragraphs (e.g. "high-quality," "proficient," "verify") are somewhat vague and certainly open to debate. What constitutes a "high-quality" assessment? How do we know whether assessments are aligned and consistent with recognized professional standards? What precisely does it mean to use an assessment in a valid and reliable manner? What is academic proficiency? What constitutes verification of a state's assessment results? Can the NAEP results be used to do this?[6]

---

[1] This is not by any means a comprehensive portrait of the accountability portion of the legislation. For instance, the legislation also specifies intermediate goals, including statewide annual measurable objectives to meet this long-term objective. Public Law 107-110, Title I, Part A, Subpart 1, Section 1111(b)(2)(H).

[2] These subgroups include racial, ethnic, and economic groups, as well as students with disabilities and those with Limited English Proficiency.

[3] Twelve years from the end of the 2001-2002 school year would be beyond a second term of the Bush administration so policy priorities may change before this deadline. The legislation specifies intermediate goals for meeting this objective. These include each state establishing "statewide annual measurable objectives" that indicate a "single minimum percentage of students who are required to meet or exceed the proficient level on the academic assessments." These minimum percentages apply separately to each subgroup of students and not all subgroups must make adequate yearly progress each year. See Public Law 107-110, Title I, Part A, Subpart 1, section 1111(b)(2)(F) through (I).

[4] Reconstitution of a school refers to the re-evaluation of all personnel staffing positions at that school.

These are certainly all important questions that create considerable disagreement among policymakers and academics. The vagueness associated with many of the provisions in the ESEA may lead to educational progress by allowing for wise policymaking as states and the federal government work together to craft policies that best fit specific local contexts. But it is also possible that this vagueness will work to the detriment of education as states, localities, and schools game accountability systems so as to best demonstrate that adequate yearly progress is being achieved.

Ways to Misrepresent Educational Realities

In recent years, standards-based reform and accountability has become a central component of school reform initiatives in most states. Virtually all states now have developed academic standards that students are expected to meet and tests to judge school and student performance against those standards.[7] In theory this guarantees that state officials, as well as the public at large, know how much students in the state are learning. But, there are a number of ways for school districts, schools, and teachers to make it appear that their students are learning more than they actually are. The most direct is outright cheating on state assessments, a method that has been used in the past on a number of occasions.[8] Other subtle (and legal) methods may also be used to either achieve or show educational gains that are not as large as they may appear on first blush. These fall under several general headings: *strategic allocation of teacher effort; the shaping of the tested pool; the makeup of a school*; *"adjustments" of states' standards*; and *tallying methods used to measure progress.*

---

[5] This was in part because of a study (Kane, Staiger, and Geppert, 2001) showing that an overwhelming number of elementary schools in North Carolina, a state widely regarded as having a sophisticated accountability system that has resulted in improved student outcomes (Grissmer and Flanagan, 1998), would have been judged as failing based on some of the originally-proposed AYP standards.

[6] This question is addressed elsewhere in this report.

[7] A number of states also are attaching "high-stakes" to these exams (Education Week, 2002).

Strategic Allocation of Teacher Effort

Probably the most common critique of accountability systems that are based on student performance on standardized tests is that they create incentives for teachers to focus their efforts on the assessments for which they (or their schools) will be held accountable. In common parlance, they will "teach to the test." Though it is common to refer to this practice with a negative connotation attached, the practice is clearly not in and of itself a bad thing.[9] Teaching to a "good" test would be quite beneficial were it to encourage teachers to focus on class material that is educationally beneficial to their students. Thus, the accompanied implicit assumption is that teaching to a test causes teachers to focus on topics deemed to be educationally unimportant for students in the long-run.[10] The curriculum itself is often said to become "narrowed" so as to focus only on tested material. For example, teachers may focus their efforts on tested subjects, such as math and English, at the expense of subjects that are not tested, such as science, a subject that is not required to be tested until 2007-08. Teachers may also spend their time simply teaching test taking skills (Education Week, 2001; Koretz et al., 1998; Schrag, 2000). Some research does suggest that accountability systems have led some teachers to incorporate standardized test content and test-taking skills into the curriculum at the expense of other material judged by many to be more educationally important (Education Week, 2001; Linn, 2000).

---

[8] For example, in May 2001 a Maryland middle school suspended seven employees for suspected cheating on state exams (Slobogin, 2001). In 1999, a cheating scandal affected teachers in schools across New York City, while in 2000 Michigan elementary and middle schools were suspected of cheating on state exams (Hoff, 1999; Keller, 2001).

[9] See, for example, Yeh, 2001.

[10] Emerging research on states with high-stakes testing regimes, such as Texas and North Carolina, suggests that states' accountability systems are having positive effects on students' achievement (Grissmer et al. 1998). The evidence connecting accountability systems to improved student performance is not, however, conclusive (Haney, 2001).

Another way teachers might strategically allocate their efforts is by focusing on only certain types of students (Elmore et al., 1996; Heubert et al., 1998). The new ESEA legislation requires the use of a system, already in place in many states, whereby schools' performance is judged based on the percentage of students who reach established benchmarks for proficiency.[11] Under such a system, it is the *pass rate* that matters for school performance, so schools have an explicit incentive to push as many students as possible beyond the point where they are judged to be proficient. This means that schools do not get credit for learning by students who are already above the proficiency level, nor do they get credit for learning by students who fail to jump the bar. Thus the system encourages a focus on those students who are just below the benchmark. Students far below the benchmark may be seen by teachers as "lost causes," and therefore not a good place to focus efforts. Research on the accountability system employed in Kentucky lends credence to this concern. It suggests that teachers have focused efforts on average or higher-achieving students to the detriment of lower-achieving students.

Shaping of the Tested Pool

One of the best ways for schools to influence accountability results is to shape which groups of students take a test. In general, the higher the percentage of students who sit for an exam, the lower the average score on that exam (or alternatively, the lower the pass rate on the exam). This is because the highest achieving students are the ones who are most likely to sit for exams on any given day. This is the reason many states require a certain percentage of students to be tested for a school to qualify for exemplary accountability ratings, and why some states explicitly factor in attendance on the day of the test when judging a school's performance (Education Week, 2001). There are, however, a number of ways that states can strategically manipulate the tested pool without showing lower attendance rates.

In the past, one way schools could manipulate their scores was by placing students into non-tested categories, such as Special Education and English Language

---

[11] Texas, for example, has an Accountability Rating System, which is based on the percentage of students in the total population and certain subgroups who reach a established benchmarks on the state assessment (the TAAS). In order for a school to receive a "recognized" rating in Texas, at least 80% of the total students and each student subgroup must pass each TAAS subject test.

Learners (ELL).[12] Such categories are sometimes exempt from testing and have mainly been exempt from counting toward schools' accountability ratings. The 2001 reauthorization of ESEA explicitly requires states to assess the achievement of students with disabilities and limited English proficiency and it requires *all* students to reach proficient levels after 12 years. This may lead to a greater focus on disabled and ELL students. One wonders, however, how exactly those provisions will work. The explicit requirement that these special classes of students be included in the accountability system goes beyond the provisions of the 1994 law that required states' standards and assessments apply to *all* students, including special education students and ELLs. Many states sought and obtained waivers from these requirements or ignored them altogether (Taylor, 2002). Even if there is strict enforcement of the 2002 law, one still might argue that incentives exist for classification of students into these special categories since students with special needs are sometimes provided with testing accommodations.

Another way that schools may influence their testing pools is through promotion and retention policies. The new emphasis on accountability is likely to encourage schools to adopt even more stringent promotion and retention policies to ensure that students are not promoted to grades where they will perform poorly on state assessments and hurt the performance of the school. Schools, for instance, may be less likely to promote students with weaker academic skills into 3rd grade, which is the first grade with required testing.[13] This is not necessarily a negative consequence of the outcome since the jury is still out on the net impact of retention on students' ultimate outcomes.[14] The research consensus, however, is that retention increases the probability of students dropping out of high school (Holmes, 1989; Grissom and Shepard 1989). Haney (2001), for instance, finds that when an exit exam in Texas was first implemented, dropout rates increased substantially, especially for African-American and Hispanic students.

---

[12] Research on the classification of students into special education categories suggests that teacher referrals for special education services are many times improperly based on student characteristics such as race, gender, and socio-economic status, rather than on a student's actual need for special services (Ortiz, 1992; Singhal, 1999; Artiles, 1994). There is little evidence on the factors influencing the classification of students into ELL status.

[13] Alternatively, they may hustle students with strong academic skills into the 3rd grade.

[14] Far more studies argue against retention than for it (Holmes, 1989), though some studies show positive academic benefits (Kerzner, 1982; Pierson and Connell, 1992; Karweit, 1999; Eide and Showalter, 2000).

The Makeup of a School

Up to this point, I have implicitly treated what constitutes a school as a given and focused on the shaping of the pool of students within schools. There are, however, some interesting ways in which some school districts or states might manipulate the definition of a school so as to make it appear that the "school" is making AYP. For example, school systems could define "schools" in such a way that they consist of specific grades or classrooms within a single building. School systems could also classify multiple distinct "school" buildings into what would be considered by states as "single" schools.[15] Thus, local school systems could, through aggregation and reclassification of "schools," have high-achieving students offset the poor performance of lower-achievers.

One can make essentially the same case for the drawing of school district boundaries. Through educational gerrymandering neighborhoods could, for instance, be carved up so students are grouped together to maximize the probability that the largest number of schools demonstrate AYP.[16] Virginia's accountability system provides an excellent example of the potential for this type of manipulation. The unit of analysis in the Virginia accountability system is the *school*, not the *students* in the school. Thus, schools in the state may move in and out of accredited status simply based on the catchment areas of those schools. In other words, an accredited school one year could be unaccredited the next because different (lower-achieving) students are redistricted into a particular school building and this clearly is not related to the performance of personnel within the school.

"Adjustments" of States' Standards: A Race to the Bottom?

The re-authorized ESEA mandates that all states establish proficiency levels that all students in the state meet or exceed by 2013-14, but, as I mention above, it is not specific about what constitutes proficiency or how this should be measured. The language in the legislation mandates that state assessments conform to "*recognized* professional and technical standards," but an examination of various state assessments used today suggests that there are in fact no universally held views about what constitutes "good" standards.[17] In fact, various groups rate states' standards quite differently in some cases. For example, under *Education Week*'s Standards and Accountability ratings, Kentucky receives an A- but the Fordham Foundation rates Kentucky as having "Trouble Ahead," meaning strong accountability attached to bad standards. Furthermore, there exists today a surprising amount of variation among states in how they rate the performance of their students in Title I (lower income) schools (U.S. Department of Education, 2001). For example, in Georgia 59 percent of Title I schools were identified as being in need of improvement while Tennessee identified only 2 percent of its Title I schools.[18] Were states to set the bar low enough, 100 percent of their students could be judged as proficient today.

---

[15] States receive student achievement information based on school codes. There is nothing that precludes states from allowing districts, for example, to specify two "school" buildings from opposite ends of a county as having the same code. From a state's perspective, this would then *de facto* be the same school.

[16] This would not work indefinitely because, holding the true achievement levels of students constant, there are only so many ways that high- and low-achieving students can be grouped to show AYP over time.

Tallying Methods Used for Measuring Progress

The reauthorization of the ESEA is also silent on the precise methodology that states should use to measure or tally progress toward meeting the goals outlined in the legislation. The specific attributes of accountability systems differ significantly between states. For instance, among the states that use tests, there is variation in the type of exam used to measure student achievement. Some use assessments developed by the state (e.g., TAAS), while others use norm-referenced tests (NRT) such as the Stanford-9. Still others employ criterion-referenced tests (CRT), such as the Terra Nova. Many states use a combination of these options. States may use different tests from one year to the next, and these may not be designed to be directly comparable from year to year. The reason is that NRTs show how students in a particular grade compare relative to other students at a particular grade level, while CRTs show the extent to which students have mastered particular skills. It is possible for students in a particular state to improve their performance on CRTs while they perform less well on NRTs (or vice versa), particularly if states adopt different standards. This combination would reflect students who are gaining proficiency on their state's standards but who are not performing as well *relative* to other students (often nationally) on the items on the NRT (which may not be closely aligned with their particular state's standards). The result of using very different types of assessments is that it would be necessary to use some secondary method to determine academic growth from year to year and thus comply with the AYP mandate. This, of course, is not a trivial or uncontroversial task.

---

[17] Public Law 107-110, Title I, Part A, Subpart I, Section 1111(b)(3)(C)(iii).
[18] U.S. Department of Education, 2001.

There are also major differences in the tallying methodologies used to assess school performance. Today states use a variety of accountability standards, such as the average scores by grade level, the percentage of students who reach established benchmarks, changes over time in these measures, and various "value-added measures" such as the school-level average of gains for individual students.[19] Some are far better than others at identifying the actual contributions of teachers and schools.[20] But regardless of the system employed, it is common to observe the so called "saw-tooth effect" — the finding that that test scores increase substantially during the initial years of a test's administration due simply to increased familiarity with the assessments, and then level off (Heubert, 1998; Koretz, 1988; Linn, 2000; Schrag, 2000).

If test scores do increase substantially during the initial years of their administration and then level off, states might introduce new assessments once they have reached the leveling off point. States may also simply change the rules of the tallying system. In Virginia, for instance, starting in 2001, the state changed the methodology used to determine schools' performance on the Standards of Learning (SOL) test, the state's assessment. The difference between the scores under the old and new methodology is that the new scores account for the performance of students who had previously failed to reach proficiency levels but had been through a remediation program and retaken the test. These students, however, are only accounted for in the numerator. This adjustment to the accountability system in Virginia has created the strange situation where, at least in theory, schools can have adjusted SOL pass rates of over 100 percent even if the majority of students at a particular grade level were not judged to be proficient. This new method of calculating pass rates also makes it appear as if the state is making greater progress towards the goal of all students in the state achieving academic proficiency.

---

[19] Other value-added measures include comparing differences between actual and regression-generated predicted scores.

[20] For instance, in my opinion, it is necessary to use a value-added methodology and account for family, student, and background factors to effectively isolate the contributions of schools and teachers. Additionally, most standardized achievement tests are designed to provide relative scores and they may be inadequate at measuring whether students have mastered particular standards (Popham, 2002).

There may well be valid reasons for Virginia altering their method to assess schools, however, it illustrates the point that such systems can be manipulated simply for the sake of changing perceived progress. The bottom line is that accountability systems may be gamed to show student achievement gains. This is possible because states have the flexibility to set their own standards, administer their own tests, and craft systems to judge student performance. Thus, one could imagine a worst case scenario where the pressure, political and otherwise, to show that students are making academic gains could create a race to the bottom in terms of standards and accountability systems.

Conclusions: Checks on the Gaming of the System?

What is to prevent states from setting low standards or the manipulating of the system of the sort described above. In theory the highly regarded national proficiency test administered in grades 4 and 8—the National Assessment of Educational Progress (NAEP) can be used to verify the reported state gains in academic proficiency. Serious manipulation of a state's system might be detected by discrepancies between state reports of students' AYP (based on state assessments) and their performance on NAEP. But, for a variety of reasons, there is considerable doubt as to whether NAEP is up to this task. One can easily imagine situations where states truly show remarkable student gains on the state assessment, but have their NAEP scores remain flat. This can occur, for instance, if a state opts to adopt standards that are not well-aligned with what is tested on the NAEP. Recent studies, in fact, have found a number of cases where states with large improvements in state test scores experienced little improvement on the NAEP (Klein et al., 2000; Koretz et al., 1998).

Discrepancies between state assessment and NAEP results would, of course, not preclude state officials from making the argument that their students are in fact gaining academically. Disputes over differences between NAEP and state assessment results will no doubt create a windfall for statisticians and testing experts in the business of equating different tests — this may be particularly difficult if many students opt out of taking the NAEP test, as they are allowed to do. The truth about student achievement will be out there, but policymakers and much of the public likely will not know what to make of the arcane statistical arguments.

A second potential check on states gaming the system is the requirement that states' educational plans be approved by the Department of Education. But, the legislation also limits the Secretary's authority by explicitly stating that the Secretary "shall not have the authority to require a State, as a condition of approval of the State plan, to include in, or delete from, such plan one or more specific elements of the State's academic content standards or to use specific academic assessment instruments or items."[21] Furthermore, unlike the provisions in an earlier proposed version of the legislation, the Secretary does not have the authority to withhold educational funding from states that are not seen to be making AYP based on the NAEP. Thus, in some respects, the Secretary of Education wields a relatively soft stick. The bottom line is that political realities will likely place some major constraints on the ability of the Secretary to influence states' educational plans. As Toch (2001) notes, there has been far less than full adoption of the testing requirements that were put in place in the 1994 reauthorization of the ESEA.

The law takes what appears to be a firm stand that *all* students be proficient in 12 years, but this is an eternity in political terms. In the meanwhile, there exists a great deal of room to make it look like real progress is being made while the reality is otherwise. It would be truly unfortunate if manipulation of the sort described above actually occurred because it would reduce the likelihood that the goals of the legislation are realized and likely serve to undermine, in the eyes of the public, the notion that standards and accountability systems can be used as a means of improving education.

---

[21] Public Law 107-110, Title I, Part A, Subpart I, Section 1111(e)(1)(F).

References

Artiles, Alfredo (1994). "Overrepresentation of Minority Students in Special Education: A Continuing Debate." *The Journal of Special Education.* 27(4).

Bishop, John H (1989). "Is the Test Score Decline Responsible for the Productivity Growth Decline?" *The American Economic Review.* Vol 79 (1).

Carnoy, Martin, Susanna Loeb, and Tiffany Smith (2000). "Do Higher State Test Scores in Texas Make for Better High School Outcomes?" Stanford University: School of Education.

Education Week (2001). Quality Counts, 2001: A Better Balance—Standards, Tests, and the Tools to Succeed. *Education Week* 20(17).

Eide, E.R. and Showalter, M.H. (2000). "The Effect of Grade Retention on Educational and Labor Market Outcomes." Forthcoming in *Economics of Education Review.*

Elmore, Richard, Charles Abelmann, and Susan Fuhrman (1996). "The New Accountability in State Education Reform: From Process to Performance." In Ladd, Helen (Ed.). Holding Schools Accountable: Performance-Based Reform in Education. Washington, D.C.: The Brookings Institute.

Erickson, Ronald (1998). "Special Education in an Era of School Reform: Accountability, Standards, and Assessment." Federal Resource Center.

Grissmer, David and Ann Flanagan (1998). "Exploring Rapid Achievement Gains in North Carolina and Texas." Washington, D.C.: National Education Goals Panel.

Grissom, J.B., and Shepard, L.A. (1989), Repeating and dropping out of school. In L.A. Shepard and M.L. Smith (Eds.), *Flunking Grades: Research and Policies on Retention.* London: Falmer, pp. 34-63.

Haney, Walt (2001). "Revisiting the Myth of the Texas Miracle in Education: Lessons about Dropout Research and Dropout Prevention." Paper prepared for the Dropout Research: Accurate Counts and Positive Interventions Conference: Cambridge, MA.

Heubert, Jay and Robert Hauser (Eds) (1998). "High Stakes: Testing for Tracking, Promotion, and Graduation." Committee on Appropriate Test Use, National Research Council.

Hoff, David J. (1999). "N.Y.C. Probe Levels Test-Cheating Charges." *Education Week* (web edition). 15 December 1999.

Holmes, C.T. (1989). "Grade level retention effects: a meta-analysis of research Studies." In L.A. Shepard and M.L. Smith (Eds.), *Flunking Grades: Research and Policies on Retention*. London: Falmer, pp. 16-33.

Kane, Thomas, Douglas Staiger and Jeffrey Geppert (2001) "Assessing the Definition of 'Adequate Yearly Progress' in the House and Senate Education Bills." Unpublished manuscript, 2001. Available: http://www.dartmouth.edu/~dstaiger/WP.html

Karweit, N.L. (1999). "Grade Retention- Prevalence, Timing, and Effects." CRESPAR Report No. 33. Web address: http://scov.csos.jhu.edu/crespar/reports/report33chapt1.html.

Keller, Bess (2001). "Dozens of Mich. Schools Under Suspicion for Cheating." *Education Week* (web edition). June 20, 2001.

Kerzner, R.L. (1982). *The Effect of Retention on Achievement*. Union, NJ: Kean College of New Jersey.

Klein, Stephen P., Laura S. Hamilton, Daniel F. McCaffrey, Brian M. Stecher (2000). "What Do Test Scores in Texas Tell Us?" RAND.

Koretz, Daniel and Sheila Barron (1998). "The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS)". RAND.

Linn, Robert (2000). "Assessments and Accountability." *Educational Researcher* 29:2.

Linton, Thomas (2000). "High Stakes Testing in Texas: An Analysis of the Impact of Including Special Education Students in the Texas Academic Excellence Indicator System." Paper presented at the Texas Assessment Conference, Austin, Texas.

National Center on Educational Outcomes (1999). "A Report on State Activities at the End of the Century." Available on the web: http://www.coled.umn.edu/NCEO/

Ortiz, Alba (1992). "Assessing Appropriate and Inappropriate Referral Systems for LEP Special Education Students." Proceedings of the Second National Research Symposium on Limited English Proficient Student Issues: Focus on Evaluation and Measurement.

Pierson, L.H., and Connell, J.P. (1992). "Effect of Grade Retention on Self-system Processes, School Engagement, and Academic Performance." *Journal of Educational Psychology* 84, 300-307.

Popham, James W (2002). "Right Task Wrong Tool." *American School Board Journal.* 189(2), 18-22.

Robelen, Erik W. (2001). "States Sluggish on Execution of 1994 ESEA." *Education Week* (web edition). 28 November 2001.

Schrag, Peter (2000). "Too Good To Be True." *The American Prospect.*

Singhal, Rebecca (1999). "Revisiting Segregation: The Overrepresentation of Minority Students in Special Education." ERIC document.

Slobogin, Kathy (2001). "Cheating scandals test schools." 23 July 2001. www.cnn.com.

Taylor, William L., and Dianne M. Piche (2002). "Will New School Law Really Help." *USA Today.* 9 Jan 2002. A13.

Toch, Tomas (2001). "Bush's Big Test: The President's Education Bill is a Disaster in the Making. Here's How He Can Fix It." *The Washington Monthly*, November.

U.S. Department of Education (2001). "High Standards for All Students." January, 2001.

Wolf, Patrick and Bryan Hassel (2000). "Effectiveness and Accountability in Special Education, Part I: The Compliance Model." Thomas B. Fordham Foundation and Progressive Policy Institute.

Yeh, Stuart S. (2001). "Tests Worth Teaching To: Constructing State-Mandated Tests that Emphasize Critical Thinking." *Educational Researcher.* Vol. 30(9). pp. 12-17.

Comments

Abigail Thernstrom

Of course I do applaud the commitment to standards and accountability that this statute represents--the effort to institute a culture of high expectations and adult responsibility, as Michael Cohen puts it.

But there are (just for starters) huge definitional problems built into the legislation--a "squishiness in terms," as Dan Goldhaber put it. For instance, in Massachusetts we already have a definition of "Proficiency." It's one that roughly corresponds to the NAEP definition, and it's a goal that 100 percent of students cannot possibly reach, even with a dozen years in which to do so. In fact, frankly, it's a ludicrous goal. At the moment, we in Massachusetts are just hoping that a respectable percentage of our kids manage to get into a low expectations, minimum skills category called "Needs Improvement." Those who do so will get a high school diploma.

In order for all students to become Proficient on the NAEP state assessments or MCAS in Massachusetts, we will have to define proficiency, which is supposed to be the goal, way, way down. All students means members of every racial and ethnic group, and I am particularly concerned about the scores of black and Hispanic kids who indeed have been subject to a "soft bigotry of low expectations," but who have so far to go. It's a picture that makes you want to cry; it's a national crime. But the problem won't be easily or quickly fixed.

In the most recent NAEP math and reading assessments, only 32 percent of all American students scored at the Proficient or Advanced level in 4th grade reading. For blacks the figure was just 12 percent, and for Hispanics 16 percent. The scores in math were even worse, with only 5 percent of blacks, 10 percent of Hispanics, and 26 percent of all students rated Proficient or Advanced in 4th grade. By 12th grade, even fewer non-Asian minority students fell into the top NAEP categories.

So let's not kid ourselves. Getting all of our students to anything close to what NAEP defines as Proficient is just not possible. It's not possible in Massachusetts or in any other state.

Diane Ravitch tells me that indeed no one expects "proficiency" to mean Proficiency by the current NAEP standard. Proficiency will mean Basic, as NAEP now defines it, although why the legislation confusingly refers to proficiency then becomes a mystery. But, in any case, getting everyone up to that Basic level is utopian enough.

Looking again at recent NAEP assessments, in 8th and 12th grade reading, a quarter of all American students have academic skills and knowledge below the Basic level. On other assessments (math in all grades and reading in 4th grade), roughly a third of the students are below Basic. And for the subgroups that are the focus of the legislation, the picture is far more dismal.

For instance, at the end of 12 years of education, 7 out of 10 African Americans lack even a Basic command of math. That is the case for more than half of Hispanics too. Reading scores are somewhat better, but still terrible. The picture for low-income children is also appalling. Seventy-three percent of kids eligible for free or partially reduced lunch are Below Basic in 12th grade math.

Sadly, even these figures are too optimistic, since not all special education and LEP students were tested.

What about trends in recent years? The news in that respect is also abysmal. We have put money and effort into reforming education with almost no returns. We are expecting this legislation to usher in a new era of steady progress. Recent trends do not suggest much reason for optimism.

Okay, these are national figures, one might say. There is considerable variation between states, surely. No Child Left Behind was largely based on the Texas model, celebrated as a story of educational success. In fact, the picture seems extremely mixed: progress in some respects, but no evidence that the state has a formula for eliminating or dramatically narrowing the racial gap in achievement. When gains are measured by scores on TAAS, they are impressive. But when TAAS results are compared with performance on NAEP assessments, the news is much less encouraging.

Thus, looking at the limited reading and math trend data for Texas NAEP during the period of educational reform, we find that the black-white gap narrowed only in 4th grade math, while the Hispanic-white difference shrunk only in math in 8th grade. In 4th-grade reading and in both science and math in the 8th-grade, the black-white gap in Texas did not change and remained very large, just as large as in the nation as a whole. Moreover, the black gains in Texas occurred at the same time as more modest progress was apparent in the nation as a whole.

Where we have no trend data, we can compare scores on Texas NAEP assessments with those of the nation as a whole, although with no high school data, we are left in the dark as to the most important question of all: How much have students learned by the time they finish 12th grade? In the elementary grades, we do find modest progress in narrowing the black-white gap, but it seems to have been wiped out by middle school--with the one exception of writing, about which legitimate questions can be raised. The Hispanic picture is more heartening, although the gains could reflect changes in the composition of the Texas Hispanic population.

We can also ask: How many Texas students moved from Below Basic to Basic? This is obviously another possible measure of success. Again, Hispanic scores are encouraging, but while black scores on some assessments rose in the elementary years, in the middle school years Texas gets only a D.

Finally, some comments about the state I know best: Massachusetts. We have been pouring money into education since 1993. There has been a major push for high standards and real accountability. And we cut the failure rate on MCAS roughly in half last year. But we still have a very large group of students who are in danger of not graduating from high school. Moreover, once again, the failure rate is strongly skewed by race. Nearly two thirds of the state's black high school students and seven out of ten Hispanics will not receive diplomas in 2003 unless they do better on one or both of the English and math tests.

And again, we're not asking that these kids get to "Proficiency," but only that they move into the very low "Needs Improvement" category--a goal that won't be easy to meet. Nor will it be easy to get the schools that educated them so poorly to institute an academic program that will make for real success in the future.

In theory, in consistently failing schools, parents will have the right to switch their kids to another public school. But where are they supposed to go? Your kid goes to an unsatisfactory Boston elementary school. Good luck in finding one that's better and has empty seats.

States and districts will have to provide technical assistance to schools that fail to make progress, the statute says. Great idea on paper, but Massachusetts already has an intervention program, as do many of the individual districts. None of them are worth much, in my view. Neither the state nor the districts really know how to turn schools--no less whole districts--around. The state board of education, on which I sit, can't run school systems. Especially because we can't fire half the teachers in a district, find (for instance) a whole bunch of teachers who really understand the structure of math and how to teach it, replaced the administrators with a great new team, change the rules governing salaries, insist on abandoning fuzzy math, make sure kids don't arrive in kindergarten already behind, etc., etc. etc.. Effective intervention is a fantasy. Or at least to a sobering extent it is. And so is really good tutoring for massive numbers of students.

In short, I don't know how we're going to meet the standard--even if it's NAEP's "Basic," and I don't know how we're going to have effective intervention within the public school system as it's currently structured. Of course, not all the news is equally bleak. Texas has a bit to show for its effort; across the educational landscape there are some good schools beating the demographic odds. But the goal established by this legislation is truly daunting.

Those, alas, are my pessimistic thoughts for the morning.

Thanks.

About the Contributors

Michael D. Casserly (Council of the Great City Schools)
Michael D. Casserly is the Executive Director of the Council of the Great City Schools. He was the organization's chief lobbyist and research director for 15 years before being named to head the group in 1992. He is an urban advocate for high standards, strong management, and adequate school funding.

Michael Cohen (Aspen Institute, former Assistant Secretary, U.S. Department of Education)
Michael Cohen is a senior fellow at The Aspen Institute. He served in the Clinton Administration in several senior policy making capacities, including Assistant Secretary for Elementary and Secondary Education and special assistant to the President for Education Policy. Previously, he served as the education program director for the National Governors' Association, and in senior positions at the National Center on Education and the Economy and the National Association of State Boards of Education.

David Figlio (University of Florida and National Bureau of Economic Research)
David Figlio, Walter Matherly Professor of Economics at the University of Florida, received his Ph.D. in Economics from the University of Wisconsin-Madison in 1995. His research on education policy has been published in top economics journals, and his work on school accountability has been funded by several federal agencies and numerous private foundations. He is currently advising the governments of Chile, Sweden and Tanzania on school policy.

Chester E. Finn, Jr. (Thomas B. Fordham Foundation)
Chester E. Finn, Jr., has devoted most of his career to improving education in the United States. A former Assistant Secretary for Research and Improvement at the U.S. Department of Education, Finn is currently John M. Olin Fellow at the Manhattan Institute and President of the Thomas B. Fordham Foundation. He is the author of 13 books and over 300 articles on education policy.

Matthew Gandal (Achieve, Inc.)
Matthew Gandal is executive vice president of Achieve, a national non-profit that helps states raise standards and achievement in their schools. Before joining Achieve, Matthew was assistant director for educational issues at the American Federation of Teachers (AFT), where he helped launch a variety of programs and publications, including *Making Standards Matter*, an annual report evaluating the quality of the academic standards, assessments and accountability policies in the 50 states.

Dan Goldhaber (Urban Institute)
Dan Goldhaber currently serves as a Senior Research Associate at the Urban Institute's Education Policy Center and is an adjunct faculty member at the Georgetown University Public Policy Institute. His research focuses on issues of educational productivity and reform at the K-12 level and the relationship between teacher labor markets and teacher

quality. He also served as an elected member of the Alexandria City School Board from 1997-2002.

Paul A. Herdman (New American Schools)
Paul A. Herdman is Director of Accountability and Evaluation Services at New American Schools, where he provides overall leadership for such services provided to states, districts, and charter schools. He has contributed to several publications on charter schools and urban education reform throughout the U.S. and has served as a public school teacher and administrator.

Brian J. Jones (Education Leaders Council)
Brian Jones is VP for Communications and Policy at the Education Leaders Council. He was formerly Associate Superintendent for Federal Programs and Policy with the Arizona Department of Education, overseeing the agency's vast federal budget and working to promote Arizona's reform initiatives at the national level. Prior to his service in Arizona, Brian spent eight years working in the United States Senate as a legislative analyst and policy advisor for Senator Pete V. Domenici (R-NM) and for the Senate Committee on Labor and Human Resources under Chairman Jim Jeffords (I-VT).

Lisa Graham Keegan (Education Leaders Council)
Lisa Graham Keegan is Chief Executive Officer of the Education Leaders Council, an organization of reform-minded education chiefs she helped found in 1995. Prior to this position she was Arizona's Superintendent of Public Instruction. Known for her focus on educational improvement and reform, including efforts for student-centered funding, charter schools, expanded school choice and an emphasis on marketplace incentives, Keegan originally developed her policies during her service in Arizona's House of Representatives, where she chaired the Education Committee and authored much of the reform legislation she now oversees.

Billie J. Orr (Education Leaders Council)
Billie J. Orr serves as President of the Washington-based Education Leaders Council. A former public school teacher and principal, Orr was appointed Arizona's Associate Superintendent for Public Instruction under Superintendent Lisa Graham Keegan. In that role, Orr led the effort to develop and implement statewide academic standards and oversee the development of the state's criterion-referenced test. Orr holds a Doctor of Education -- as well as an M.A in Reading and a B.A. in Education – from Arizona State University in Tempe, AZ.

Mark D. Reckase (Michigan State University)
Mark D. Reckase is a professor of Measurement and Quantitative Methods in the College of Education at Michigan State University where he works on testing and data analysis issues. Prior to coming to MSU, he was an Assistant Vice President at ACT, Inc. where he worked on the contract to set achievement levels on NAEP.

Nelson Smith (New American Schools)
Nelson Smith is Managing Director for New School Services at New American Schools. He was the first Executive Director of the District of Columbia Public Charter School Board; was Vice President for Education and Workforce Development at the New York City Partnership; and served for six years in the US Department of Education's Office of Educational Research and Improvement.

Abigail Thernstrom (Massachusetts State Board of Education)
Abigail Thernstrom is a Senior Fellow at the Manhattan Institute, a member of the Massachusetts State Board of Education, and a commissioner on the United States Commission on Civil Rights. She co-authored *America in Black and White: One Nation, Indivisible* (Simon & Schuster) and is working on a new book: *Getting the Answers Right: The Racial Gap in Academic Achievement and How to Close It.* Her 1987 work, *Whose Votes Count? Affirmative Action and Minority Voting Rights* (Harvard University Press) won the American Bar Association's Certificate of Merit. She also writes frequently for a variety of journals and newspapers.

Richard J. Wenning (New American Schools)
Richard J. Wenning is Vice President for the Education Performance Network, the professional services arm of New American Schools. Previously, he was President of Choice Strategies Group, a consulting firm based in Washington, DC. He also served as Senior Policy Advisor to the CEO of the D.C. Public Schools during the school district's takeover. Prior to that he served as Clerk for the Senate Appropriations Subcommittee on the District of Columbia and as staff to the Subcommittee on Labor, Education, and Health and Human Services.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

NOTICE

Reproduction Basis

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").

EFF-089 (1/2003)